Trade Study of Implementation of Software Defined Radio (SDR): Fundamental Limitations and Future Prospects. Final Report

December 9, 2008

Charles W. Bostian
Jeffrey H. Reed
J. Randall Nealy
Feng Ge
Julia Mays, Editor

Wireless@Virginia Tech
Mail Code 0111 • Virginia Tech
Blacksburg, VA • 24061-0111
www.wireless.vt.edu

James Neel

Cognitive Radio Technologies, LLC
147 Mill Ridge Road • Suite 119
Lynchburg, VA • 24502
www.crtwireless.com

Sponsored by
Defense Advanced Research Projects Agency (DOD)
Strategic Technology Office
ARPA Order AF89-00
Issued by U.S. Army Aviation and Missile Command Under
Contract W31P4Q-07-C-0210

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
09-12-2008	Final Technical Report	09-11-2007 - 09-09-2008

4. TITLE AND SUBTITLE
Trade Study of Implementation of Software Defined Radio (SDR): Fundamental Limitations and Future Prospects

5a. CONTRACT NUMBER
W31P4Q-07-0210

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Charles W. Bostian, Jeffrey H. Reed, J. Randall Nealy, Feng Ge, Julia Mays, James Neel

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Wireless@Virginia Tech, Mail Code 0111, Virginia Tech, Blacksburg, VA 24061-0111

8. PERFORMING ORGANIZATION REPORT NUMBER
430304

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency (DOD)
Strategic Technology Office
3701 North Fairfax Drive
Arlington, VA 22203-1714

10. SPONSOR/MONITOR'S ACRONYM(S)
DARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Software Defined Radio (SDR) technology is commonly advocated for waveform and frequency-agile radios. It works well for simple signals and limited bandwidths, less so for complex broadband waveforms. Whether these difficulties reflect theoretical limits or design choices was unknown since few quantified limits exist. Using literature surveys and analysis this report explores fundamental limits to SDR bandwidth and waveform complexity, design tradeoffs, closeness of current technology to these limits, and future trends. For fixed front ends, SDR bandwidth is limited by analog-to-digital converter (ADC) bandwidth, dynamic range, and aperture jitter. The last dominates ADC fabrication limitations and GSM-like dynamic range for 2.5 GHz digitized bandwidths is theoretically impossible - not a fabrication limitation. Flexible front-ends are important as 2nd-order products limit practical instantaneous bandwidth to less than an octave. Increasing parallelism should improve processor performance until 2025, reaching a limit of 15nanowatts per million multiply-and-accumulate operations per second. Multicore processors will alleviate latency.

15. SUBJECT TERMS
software defined radio, radio frequency design, radio frequency systems, radio architecture, analog-to-digital converter, digital signal processor, digital execution latency

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Charles W. Bostian
U	U	U		118	19b. TELEPHONE NUMBER *(Include area code)* 540-231-5096

Table of Contents

List of Figures

List of Tables

1 Task Objectives

Software Defined Radio (SDR) technology has been widely embraced as the way to develop waveform and frequency agile radio platforms. While SDR has achieved good results when implementing simple waveforms, achievements for more complex waveforms like IEEE 802.11g have been mixed. Processor requirements can be significantly greater than anticipated, and it is often difficult to meet the stringent timing requirements of the MAC layer.

To date, it has been difficult to assess if these shortcomings are the result of fundamental limitations to the technology or could be overcome with sufficient engineering effort because there is little basis on which to judge the performance of an SDR implementation so questions persist about whether Moore's Law, faster processors, and advancing microelectronics technology will solve the current implementation problems. There are no theoretical results analogous to the Shannon limit for a communications system that could tell us how close an implementation is to the best possible. There are few quantifiable relationships between requirements and resources to guide a designer in what can be achieved with the available hardware components. Limits imposed by the laws of physics cannot readily be distinguished from implementation issues. In this report, we address these issues by answering:

(a) What are the fundamental limits to SDR performance?
(b) How close is current technology to the fundamental limits?
(c) What are the fundamental tradeoffs created by these limits?
(d) Because of these tradeoffs, which requirements are best suited to which platform(s)?
(e) How much improvement in performance can be expected in the near future?

2 Technical Problems and Methodology

The following describes technical problems of SDR and how they relate to SDR architectures, the methodologies employed, and the organization of the remainder of this document.

2.1 Overview of SDR Architectures and Technical Challenges

A SDR receiver system is easily visualized as shown in Figure 2.1 as an antenna followed by an analog signal processing chain (filters, RF amplifiers) followed by an analog-to-digital converter (ADC) followed by a digital signal processing system. Information flows from the antenna to the receiver output. A transmitter reverses the information flow and replaces the ADC with a digital-to-analog converter (DAC). Within this deceptively simple architecture there are many variations. For example, the analog signal processing system can range from nonexistent (The ADC or DAC are connected directly to the antenna.) to a cascade of filters, mixers, and amplifiers. The ADC or DAC can operate at RF, IF, or baseband.

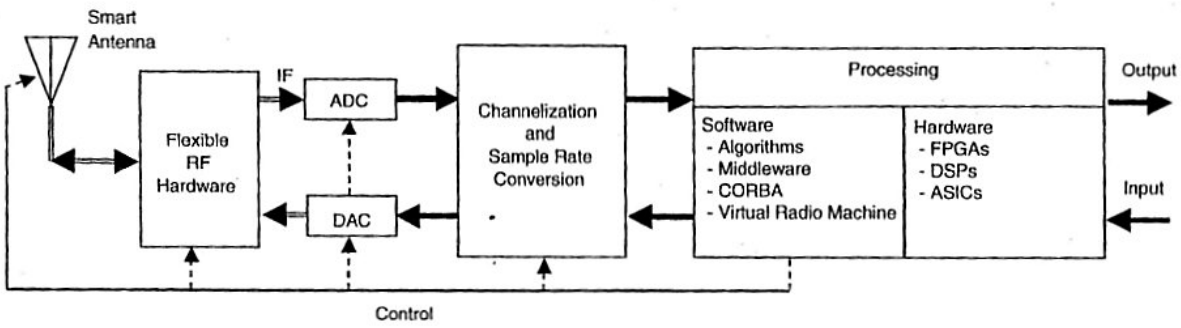


Figure 2.1 Block diagram of a generic software defined transceiver [Reed_02]

Another way to bound the architectural choices is shown by Figures 2.2 and 2.3, which contrast a direct sampling receiver with a DSP-based superheterodyne receiver.

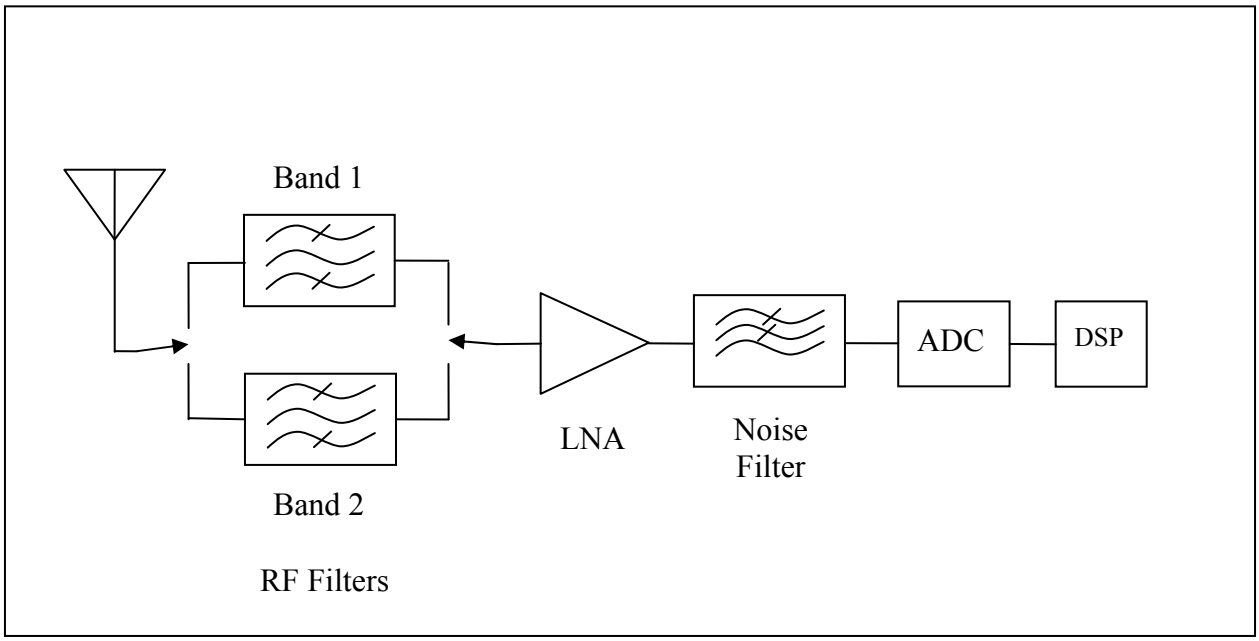


Figure 2.2 Block diagram of a direct sampling receiver

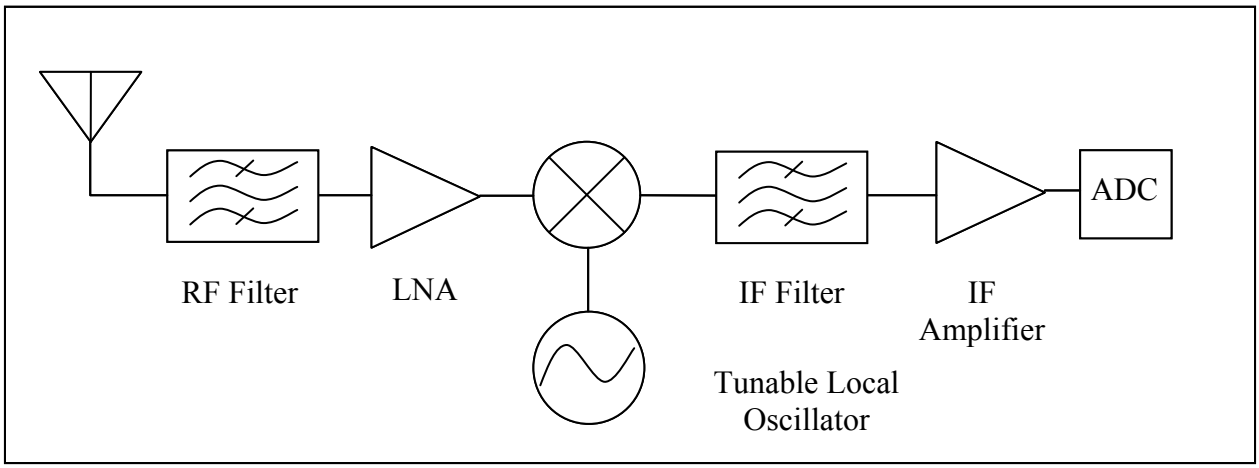


Figure 2.3 Block diagram of a DSP-based superheterodyne receiver.

In a *direct sampling receiver* (Figure 2.2), a wideband signal is digitized at RF, perhaps after some band limiting filtering and amplification. In a more realistic architecture (Figure 2.3) that we call the DSP-based superheterodyne receiver, a narrow band signal is down converted to IF before digitizing. [Perlman_08] terms this an *SDR with spectral signal processing* as in Figure 2.4. While a direct sampling architecture would allow operation over an extremely wide bandwidth (approaching the proverbial "DC to light"), practical considerations (particularly the dynamic range of the ADC) restrict designers to

the DSP-based superheterodyne layout. Efforts to provide wide bandwidth focus on providing frequency agile analog components between the antenna and the digital stages, as in the *tunable component architecture* of [Org_07A], shown in Figure 2.5.

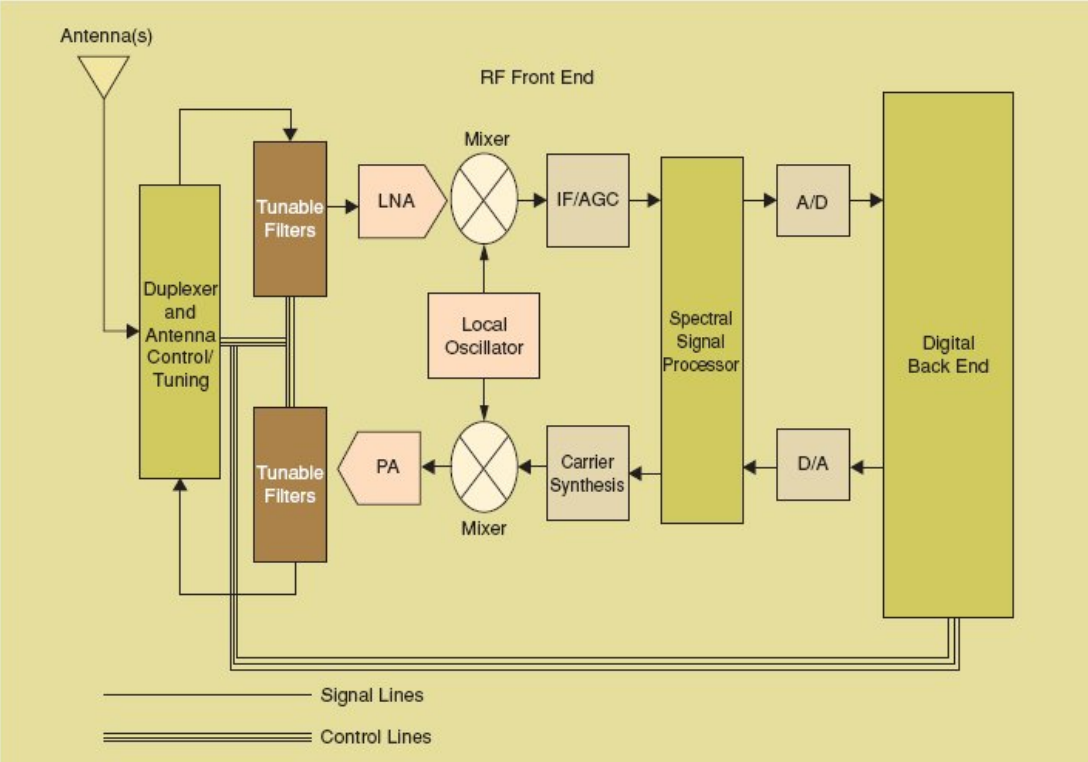


Figure 2.4 An SDR with spectral signal processing. From [Perlman_08]

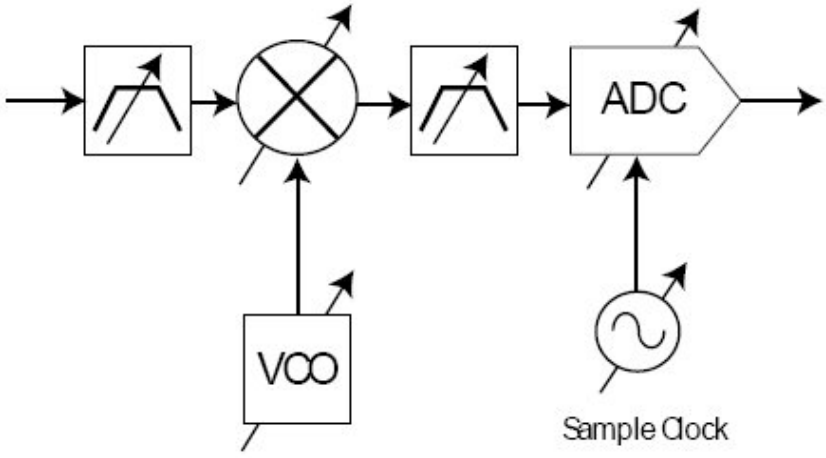


Figure 2.5. Tunable component architecture. From [Org_07A]

Figure 2.6 from [ORG_07B] shows an interesting way to look at some of the requirements for the subsystems in the tunable component architecture. The values plotted on the vertical axes represent values (maximum or minimum, as appropriate) taken from the specifications for three standard waveforms. The composite performance envelope bounded by the minima and maxima presents an interesting picture of the range of values required, while the peak values represent the specifications of a receiver that would meet the standards of all three waveforms. But it does not reflect the tradeoffs that could be made between these specifications.

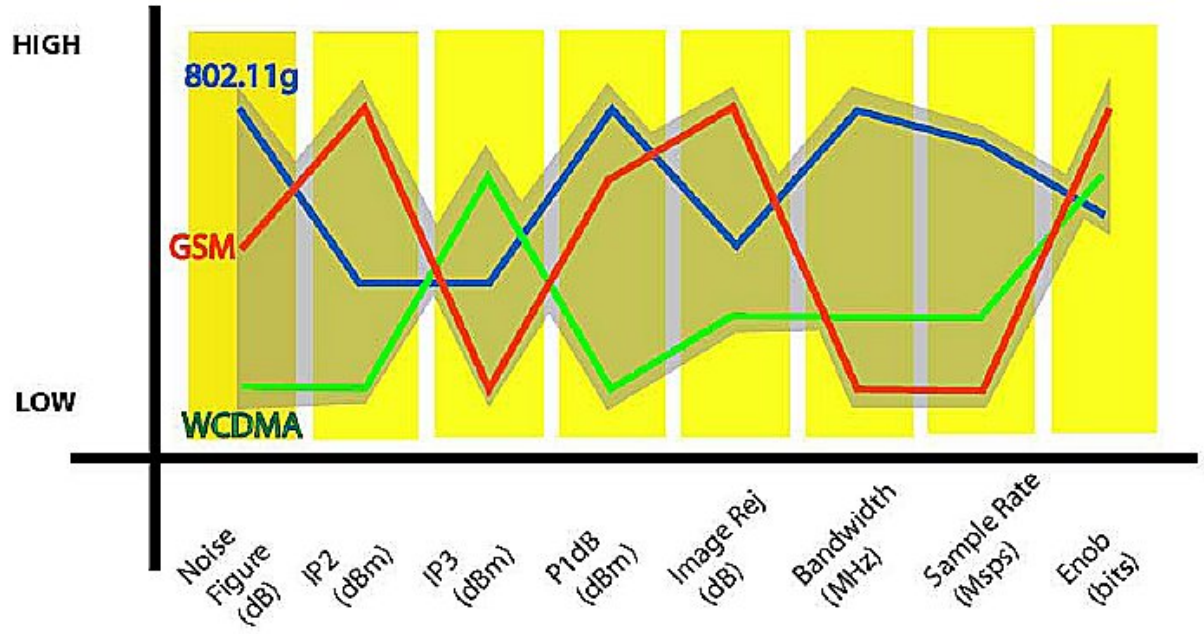


Figure 2.6 Tunable transceiver performance envelope. From [Org_07B].

2.2 Methodologies

We concluded early in the study that wide choice of architectural alternatives and the number of variables within those variations make the problem too unwieldy for a trade study without major simplifications. Accordingly, we decided to simplify the problem by beginning with the "ultimate" SDR receiver: an antenna connected to an ADC or a DAC (We will refer to the ACD and DAC as "converters.") and a digital signal processor (DSP), with almost all analog components eliminated. We then explore the operation and figures of merit of converters and seek to establish both the current practical values and theoretical limits of their performance and relate these to receiver performance. The next step is to move one step backward toward baseband and look at DSP performance. Again we define figures of merit, attempt to quantify their performance limits and relate these to receiver specifications. In both cases (converters and DSPs) we summarize the performance of devices currently on the market and attempt to identify future trends.

With the digital part of an SDR thus categorized, we look at latency issues, particularly those imposed by hardware and software architecture. Finally we look at analog RF components and antennas, identifying trends and asking how these are related to SDR performance.

To analyze trends in ADC performance, an extensive Excel database was constructed. This was built on a previous database constructed by Virginia Tech students Bin Le and Thomas Rondeau which covered data converters up to 2004. The original database contained 914 parts and the updated database contains 1523 parts. The new database removes the price entries (deemed unreliable for purposes of comparing initial price points when examining current prices at a single point of time) and adds analog bandwidth entries for all parts (including previously surveyed parts).

To analyze trends in DSP performance, an Excel database was constructed from a survey of the parts available from the following manufacturers: Texas Instruments (TI), Analog Devices (ADSP), and Freescale. To construct the DSP database, every DSP datasheet posted on each surveyed manufacturer's website was downloaded and used to collect data in the following categories: part #, clock speed, number of cores, peak # operations per cycle, peak number of MAC operations per cycle, typical core voltage, typical core current, core power consumption, bit field-width, numeric representation, year of first manufacture, and fabrication process. To augment this trend analysis, reports from Intel and the International Technology Roadmap for Semiconductors (ITRS) were reviewed to facilitate discussion of trends in general purpose processors (GPPs) and transistors.

Our analysis of latency is based on our analysis of architectural and software issues along with extensive measurements made on *GNU Radio* SDR code running on a General Purpose Processor (GPP). While the measured data represent only one particular SDR implementation, we feel that they are a good illustration of latency related problems and that we can draw some general conclusions from them.

Our study of the RF system issues had several parts. First it determined theoretical limits on receiver sensitivity that are imposed by ambient noise. The study then determined inherent limitations of antennas and addresses the difficulty of matching an antenna's frequency agility with that of an ideal SDR. The last part of the study considered the issues involved with the placement of the analog-digital boundary in the receiver chain and ways in which overall SDR performance is affected by analog components performance and their interaction with the digital components.

2.3 Document Organization

The remainder of this report is organized as follows. Section 3 compactly summarizes the key insights from this study. Section 4 and Section 5 present the results of the ADC and processors studies, respectively. Section 6 covers the results of the study into software

latency issues on SDRs and Section 7 presents the results of the study into RF system issues.

REFERENCES

[Org_07A] E. L.Org, R.J. Cyr, G. Dawe, J. Kilpatrick, and T. Counihan, "Software Defined Radio – Different Architectures for Different Applications," paper 1.4-5, SDR Forum Technical Conference 2007, November 2007.
[Org_07B] E. L.Org, R.J. Cyr, G. Dawe, J. Kilpatrick, and T. Counihan, "Design Tradeoffs in Making a Tunable Transceiver Architecture for Multi-band and Multi-mode Handsets," paper 1.3-4, SDR Forum Technical Conference 2007, November 2007.
[Perlman_08] B. Perlman, J, Laskar, and K. Lim, "Fine-Tuning Commercial and Military Radio Design," IEEE Microwave Magazine, Vol. 9, No. 4, pp. 95-106, August 2008.
[Reed_02] J.H. Reed, *Software Radio: A Modern Approach to Radio Engineering*, Prentice Hall, 2002.

3 Summary of Conclusions

While discussed in greater detail in the remainder of the report, the following highlights key conclusions from our study. Based on our inquiry into SDR architectures, we drew the following critical conclusion.

- The bandwidth of a direct sampling receiver is limited by the ADC bandwidth, dynamic range, and aperture jitter. A receiver digitizing all signals between DC and, say, 2.5 GHz is impossible, not just unrealizable given current technology limitations.

In our study of the relationships, trends, and limits to ADC performance described in greater detail in Section 4, we concluded the following:

- Improvements in data converter performance have been driven by introductions of new architectures and Moore's Law.
- ADC performance should continue to improve for the next 16 years.
- Current improvements in ADC performance are being realized as faster sampling rates with approximately the same dynamic range.
- ADC power consumption increases at approximately the square of improvements to ADC performance.
- Aperture jitter is the dominant fabrication limit to ADC performance.
- The ideal SDR architecture is physically impossible.
- Flexible RF front ends will be necessary for wide-band SDR.

In our study of the relationships, trends, and limits to processors described in greater detail in Section 5, we concluded the following:

- While the growth in GPP clock speeds have leveled-off, this has not yet happened for DSPs.
- The benefits of Moore's Law are being realized in increasing parallelism which is dramatically improving computational efficiency.
- The picoChip products are significantly ahead of traditional DSPs in terms of computational efficiency.
- Cost may become a significant limiting factor in the near future as tool costs continue to rise exponentially while revenues have flattened.
- At current rates, Moore's Law faces a physical limit in 16 years.
- Based on our transistor projections and a hypothetical "bare-bones" DSP, we estimate a limit of 15 nW per MMACS.

In our study of the relationships, trends, and limits of software and hardware architectures as presented in Section 6, we concluded the following:

- Latency problems are unavoidable because Digital Signal Processors (DSPs) and General Purpose Processors (GPPs) are based on the Von Neumann architecture which has a memory hierarchy and an operating system (OS) which introduces run time uncertainty.
- Additional latencies are introduced by the speed differences between OS and memory and between memory and I/O devices.
- The sequential signal processing inherent in GPPs introduces additional execution latency. On the other hand, ASICs, FPGAs, and analog components all process signal in a parallel/pipeline fashion (A continuous sequence of signals is executed simultaneously by a sequential set of components.).
- The latency problem can be alleviated by employing multicore processors and by developing a hybrid architecture consisting of an embedded GPP and a reconfigurable FPGA, plus some auxiliary ASICs.

Our RF systems conclusions from Section 7 are:

- Receiver sensitivity is ultimately limited by external natural and man-made noise. An optimum noise figure can be calculated based on measured noise levels. At frequencies below about 1 GHz where the external noise is high, one should design for best dynamic range given an acceptable sensitivity.
- Antenna technology is constrained by the three way tradeoff between physical size (expressed in wavelengths), bandwidth and efficiency. Antenna size reduction may result in significant inefficiency or bandwidth reduction. Small antennas may have such high Q that modulation bandwidth is restricted.
- Receiver architectures having significant filtering prior to the gain and ADC stages have a significant advantage in intermodulation performance over architectures which amplify and convert wide frequency bands.
- Receivers using fixed RF filtering must be limited to less than one octave in bandwidth or tuning range in order to avoid significant performance compromise due to second order intermodulation products.
- Contact type MEMS switches currently have hot switching ratings much too low for practical use in uncontrolled environments such as receiver antenna circuits.

4 Analog-to-Digital Converters[1]

The data converter (analog-to-digital converter – ADC – and digital-to-analog converter – DAC) significantly impacts the performance of the overall radio through factors like the radio's power consumption, dynamic range, bandwidth, and total cost. Additionally, data converters affect transceiver design as better data converter performance is needed for broadband IF sampling than for a narrowband super-heterodyne receiver, with the former more frequently used in SDR designs. Many assess how close a radio comes to an ideal software radio by the data converters' proximity to the antenna [SDRF][Mitola_95]. In this view, the ideal software radio requires the placement of the data converter next to the antenna. In this architecture, the data converter samples the RF signal and the down-conversion process is completed in the digital domain, obviating the need for nettlesome analog components. Such an architecture places some rather extreme requirements on the data converter:

- a high sampling rate to support wide signal bandwidths
- a large number of effective quantization bits to support a high dynamic range
- an operating bandwidth of several GHz to allow the conversion of a signal over a greatly varying (and theoretically arbitrary) range of frequencies
- exhibit a large spurious free dynamic range to allow for the recovery of small-scale signals in the presence of strong interferers while producing very little distortion
- minimal power consumption and cost

In general, satisfying these goals far exceed the capabilities of currently available technology. Thus priorities must be set and tradeoffs between bandwidth, dynamic range, power consumption, and cost must be made to find an acceptable design solution for not only the data converter, but also for the entire radio architecture.

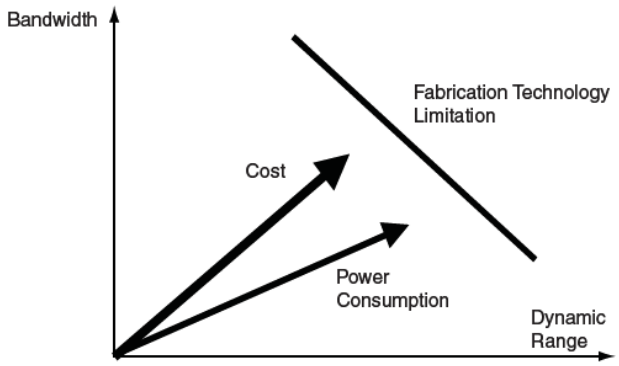


Figure 4.1: Dynamic Range, Bandwidth, Power, and Cost are fundamental ADC tradeoffs constrained by technology limitations

[1] Much of the material through Section 1.1 is taken from [Reed_02].

In light of these tradeoffs and limits, this study reports the following key findings:

- Improvements in data converter performance have been driven by introductions of new architectures and Moore's Law.
- ADC performance should continue to improve for the next 16 years.
- Current improvements in ADC performance are being realized as faster sampling rates with approximately the same dynamic range.
- ADC power consumption increases at approximately the square of improvements to ADC performance.
- Aperture jitter is the dominant fabrication limit to ADC performance.
- The ideal SDR architecture is physically impossible.
- Flexible RF front ends will be necessary for wide-band SDR.

Figure 4.2: Summary of key findings by this study.

The remainder of this section discusses how we came to these conclusions with the following structure. Section 4.1 reviews the relationships between key ADC characteristics and their relationship to SDR performance. Section 4.2 reviews the forces driving ADC design and trends in key ADC performance metrics. Section 4.3 estimates fundamental limits to ADC performance and discusses how these limits will impact SDR design and implementation.

4.1 Relationships of Key ADC Characteristics

The following reviews key ADC characteristics, their relationships with each other and their relationships with typical SDR performance metrics.

4.1.1 Overview Key ADC Characteristics

Considered abstractly, data conversion translates signals between continuous and discrete representations. An ADC changes an analog signal, which is continuous in both time and value, to a digital signal, which is discrete in both time and value. Conversely, a DAC transforms a digital signal into an analog signal. To translate signals from analog to digital representations, an ADC must perform two fundamental steps: sampling and quantization. Sampling changes a signal that exists continuously in time to a signal that is nonzero only at discrete instances of time. Quantization changes a continuous valued signal to a discrete valued signal. An ADC performs sampling followed by quantization to translate an analog signal to a digital signal. A DAC reverses this process, changing a digital signal into a continuous valued, continuous time signal through voltage mapping and interpolation.

As shown in Figure 4.3, data converters perform two fundamental operations: mapping time information and mapping amplitude information. For our purposes many of the

fundamental properties of and limits to data conversion will result from the sampling and quantization processes.

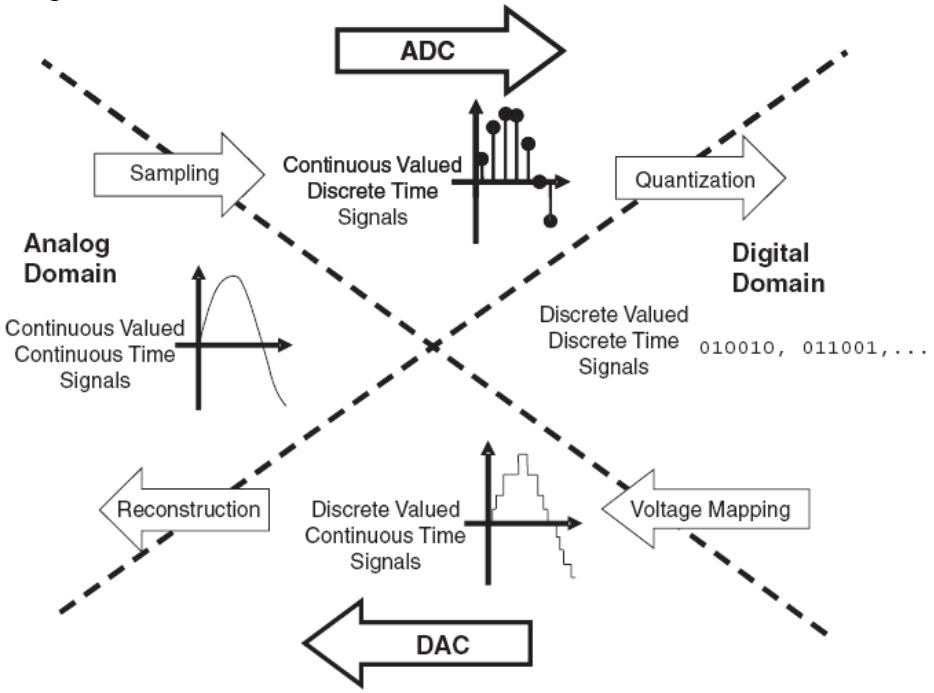


Figure 4.3: Data converters translate signal representation between continuous and discrete signal representations in both time and amplitude.

The following discusses the practical considerations associated with sampling and quantization, key ADC performance metrics of dynamic, and important practical limits to ADC performance of aperture jitter and analog bandwidth.

4.1.1.1 Fundamental ADC Process: Sampling

The sampling operation converts a signal from a continuous time representation to a discrete time form. Theoretically, sampling is a completely reversible process as long as the sampling rate exceeds twice the input signal's bandwidth – a condition concisely expressed as Nyquist's sampling theorem. In practice, the actual implementation of the sampling process places limits on the reversibility of the sampling process.

While implemented in different manners, the sampling process is frequently conceptualized as a sample-and-hold circuit similar to the one shown in Figure 4.4. When triggered by a pulse (CLK), a switch (shown as MOS transistor M_1) closes and charges a holding capacitor (C_h) with the input voltage, V_{in}. The switch then opens and the input voltage remains as the output voltage V_{out}. Variations in the timing in the sampling clock and the exact value of the effective hold capacitance significantly

influence ADC performance (clock -> aperture jitter & sampling rate; capacitance -> analog bandwidth & maximum sampling rate).

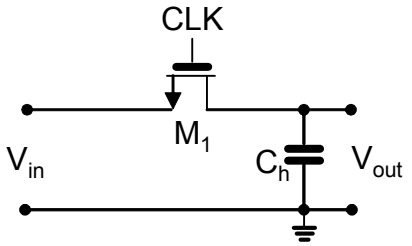


Figure 4.4: A simple sample and hold circuit.

4.1.1.2 Fundamental ADC Process: Quantization

Quantization is the process of mapping a continuous valued signal onto a discrete set of levels. In data converters, the quantization process is characterized by a few simple parameters: the number of available quantization levels, the range of quantizable input voltages, and the width of a quantization level, or *step size*. The number of levels in the discrete set is determined by the number of bits B used and is given by

$$\#\text{ quantization levels} = 2^B \tag{4.1}$$

Quantization levels are typically designed to be uniformly distributed, so it is logical to describe the data converter's resolution in terms of its step size, generally represented by the symbol Δ. Typically, the quantization levels are uniformly distributed over the range from $[-V_{FS}/2\ V_{FS}/2]$, where V_{FS} is the maximum voltage swing (or full-scale voltage) that can be input into the data converter. For a uniform distribution, the quantization step size corresponds to the value represented by the least-significant-bit (LSB) of the quantized signal and can be calculated by

$$LSB = \Delta = V_{FS} / 2^B \tag{4.2}$$

An example of a circuit used for quantization is shown in Figure 4.5 which depicts a 3-bit flash ADC architecture. Here the reference voltage is divided into 8 ($8 = 2^3$) different levels ranging from $-V_{FS}/2$ to $V_{FS}/2$ via a voltage ladder where 7 "rungs" are designed to supply a reference voltage level at the boundary of each quantization level. The analog input signal (perhaps input from a sample-and-hold circuit like the one shown in Figure 4.4) is compared against this array of reference voltage levels to determine which quantization level the input signal falls in.

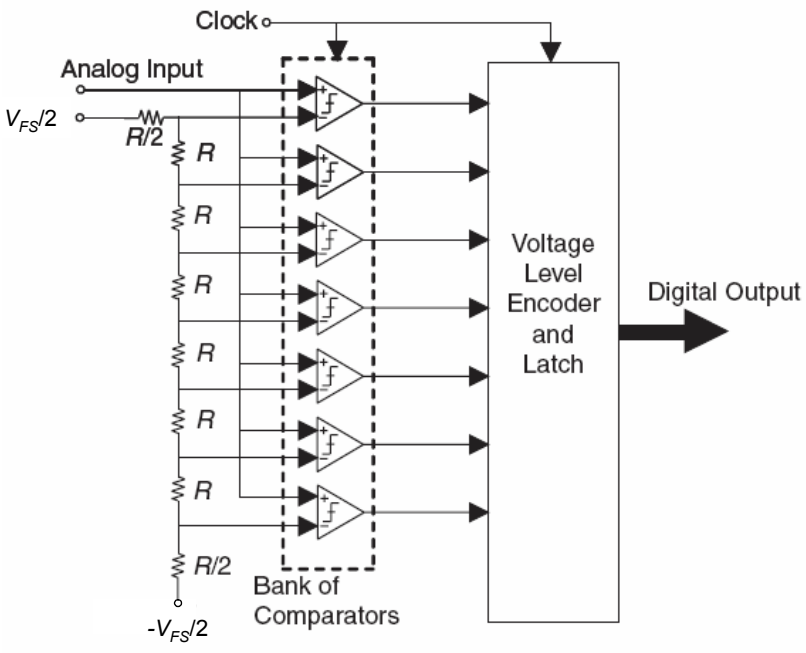


Figure 4.5: A three-bit flash ADC architecture

Since quantization approximates a continuous range of values with a discrete value, the quantization process introduces unrecoverable errors to the signal conversion process. Since these errors are generally (but not always!) uncorrelated with the signal and from sample to sample, they are frequently modeled as a white noise source termed *quantization noise*. In a typical, well-designed ADC, the maximum error that can be introduced by quantization noise is equal to one half of the largest gap between quantization levels. For an ideal data converter with a uniform distribution of quantization levels and ignoring the possibility of overranging (signals whose amplitude extend beyond $\pm V_{FS}/2$), the maximum quantization error is given by

$$\pm LSB/2 = \pm V_{FS}/2^{B+1} \text{ (V)} \quad (4.3)$$

For a sinusoid signal occupying the full range of the ADC without overranging (or clipping), the resulting signal-to-quantization ratio (SQNR) is given by:

$$SQNR = 6.02B + 1.763 \text{ (dB)} \quad (4.4)$$

A variety of techniques exist for improving the effective SQNR of an ADC (e.g., companding), but the most general is oversampling, i.e., sampling at a frequency greater than the Nyquist rate (twice the signal bandwidth) and filtering the out-of-band energy which also discards out-of-band quantization noise power. As quantization noise power is independent of the sampling rate, the power spectral density for quantization noise decreases with increasing sampling frequency and the in-band noise power decreases accordingly.

The average SQNR at the output of an oversampled ADC with a peak-to-average-power ratio of η is given by:

$$SQNR = 6.02B + 4.77 - 10\log_{10}\eta + 10\log_{10} OSR \ \text{(dB)} \qquad (4.5)$$

where the oversampling rate (OSR) is defined as OSR ≡ F_s / 2 B_s where B_S is the signal bandwidth. Note that (0.5) implies that not only does a high Peak-to-Average-Power-Ratio signal like OFDM significantly degrade power amplifier performance, it also degrades ADC performance. Also note that with proper filtering, oversampling can permit significant extension of the *dynamic range* of the data converter (see Section 4.1.1.3).

A seemingly-related phenomenon due to its terminology is *undersampling* or *bandpass sampling*. In undersampling, the Nyquist sampling rate is still satisfied (sampling at twice the bandwidth), but the highest frequency component of the sampled signal exceeds the Nyquist rate. In the digital domain, the sampling process creates a replica of the sampled signal at multiples of the sampling rate. For real signals this appears as mirror images of the sampled signal at multiples of half the sampling rate. Each region defined from [n F_s/2 to (n + 1) F_s/2] is called a *Nyquist zone* with the region defined for n=0 call the *first Nyquist zone*. Typically all signals that would fall in Nyquist zones above the first Nyquist zone are passed through a lowpass filter in the analog domain prior to digitization to reduce the amount of undesired signal energy that lies in the first Nyquist zone in the digital domain. In undersampling, the analog filter zeros out the baseband frequencies and allows energy from higher Nyquist zones to then be replicated in the first Nyquist zone with minimal interference. In this way, undersampling can be used as part of a downconversion process and eliminate an RF stages. However, the performance of an undersampling ADC is highly dependent on the performance of the analog filters, the ADC's analog bandwidth (Section 4.1.1.5) and the ADC's aperture jitter (Section 4.1.1.4).

4.1.1.3 Dynamic Range and SINAD

In wireless applications, many signals impinge upon a radio's antenna. To successfully process the desired signal, the data converter must be able to recover the desired signal in the presence of the interfering signals. The spread in signal powers between interfering and desired signals over which the desired signal can still be detected is the ADC's *dynamic range*. As SDR implementations are envisioned as using a wideband front-end which will mean capturing more interfering signal power, the dynamic range of an ADC is a very important consideration for SDR.

To appreciate the wide range of signals that a data converter must accommodate, consider the GSM-900 receiver sensitivity specification. As illustrated in Figure 4.6, GSM-900 requires the receiver be able to recover a -101 dB signal in the presence of a -13 dB interferer, thereby specifying a dynamic range requirement of 88 dB (88 = 101-13).

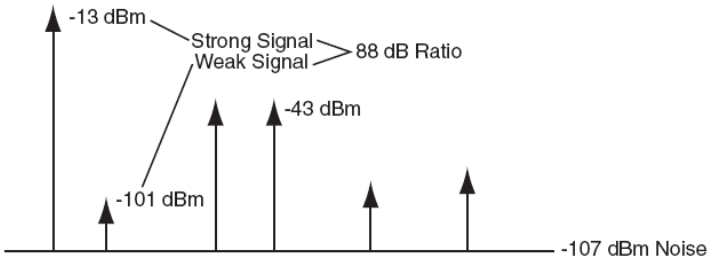


Figure 4.6: Illustration of the dynamic range requirements of GSM-900 [Brannon_96]

At a minimum, this would mean that the SQNR of the ADC should be 88 dB. In practice, the dynamic range of an ADC is never as large as the ADC's SQNR since input noise and harmonics can significantly decrease dynamic range. When these factors are taken into account, an ADC's dynamic range (DR) can be expressed as shown in (0.6)

$$DR = 10\log_{10}\left(\frac{V_{FS}^2/12\eta}{N_Q + N_T + \sum_{i=1}^{\infty} P_i}\right) \tag{4.6}$$

where N_T is noise (thermal and otherwise) input to the ADC, P_i is the power of a harmonic, and N_Q is quantization noise or

$$N_Q = V_{FS}^2/12\left(2^{2B}\right) \tag{4.7}$$

Dynamic range is frequently approximated by considering only part of the subtractive components, for example, just the distortion (harmonic) components. One measure of a data converter's relative distortion due to harmonics is its *total harmonic distortion* (THD). For data converters, THD is defined as the ratio of the sum of the powers of the harmonics, P_i, $i = 1, 2, \ldots, \infty$, which lie in the first Nyquist zone, to the power in the fundamental, P_0. THD is given by (0.8)[Telecom_96].

$$THD = 10\log_{10}\left(\frac{\sum_{i=1}^{\infty} P_i}{P_0}\right) \text{ (dB)} \tag{4.8}$$

While all harmonics present will degrade the functioning of the data converter, the analysis of how the harmonics affect the performance of a data converter can be simplified by considering the dynamic range effects of only the strongest spurious component within the first Nyquist zone. The dynamic range available in the presence of the strongest spur is the SFDR of the data converter and is calculated as the ratio of the power of the full-scale power of the desired signal to the largest harmonic (typically the third-order harmonic) as shown in (0.9).

$$SFDR = 10\log_{10}\left(\frac{P_0}{\max(P_i)}\right) \text{ (dB)} \tag{4.9}$$

Note that this differs slightly from the definition of SFDR used in receiver chain terminology where SFDR specifies the greatest power level to which the input signal can be driven before a harmonic appears above the noise floor. This difference in usage comes from ADCs typically being operated as close to their full-scale voltage range as possible to minimize the effects of quantization noise. This use of SFDR and its relationship with the data-converters SQNR can be visualized as shown in Figure 4.7.

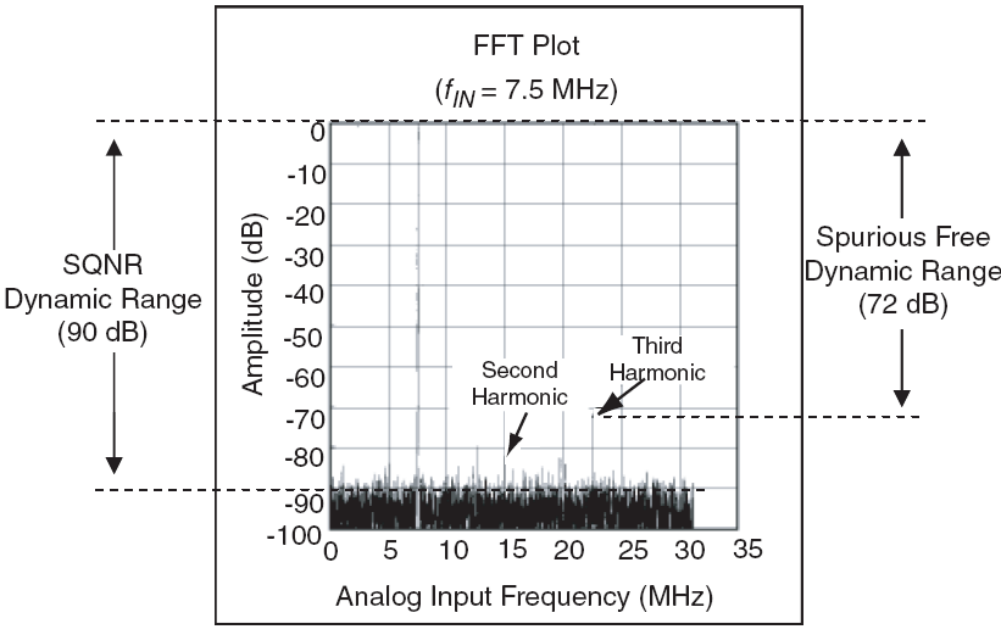


Figure 4.7: SFDR and SQNR of a Data Converter [MAX_00].

Since both noise and distortion degrade signal quality, another commonly used measurement of a data converter's performance is its *signal-to-noise-and-distortion* (SINAD) ratio. For data converters, the SINAD ratio is defined as the ratio of the signal power, P_0, to the sum of all the noise sources, N, plus the distortion from the sum of the harmonics, P_i, within the first Nyquist zone

$$SINAD = 10\log_{10}\left(\frac{P_0}{N + \sum_{i=1}^{\infty} P_i}\right) \text{ (dB)} \tag{4.10}$$

Note that (0.10) is equivalent to (0.6) when the signal is full scale and that when distortion terms are dominated by a single harmonic (most data converters), SFDR will be close to SINAD.

Frequently, SINAD is expressed in terms of the effective number of bits (ENOB) with which the ADC is operating. For sinusoidal signals, this can be calculated as shown in (0.11)

$$ENOB = (SINAD - 1.763)/6.02 \tag{4.11}$$

where SINAD is expressed in dB and 1.763 is the same offset factor (η) due to the "peakedness" of the input signal as used in (0.4) and (0.5). Because ENOB compactly captures so much of the relevant information to the performance of an ADC in a single measure that is similar to a typical ADC design metric (quantization bits), we use ENOB for data converter comparisons throughout this section.

4.1.1.4 Aperture Jitter

Due to circuitry limitations in the ADC's clock and clock distribution network, some variation in the clock timing can be expected. While it may be possible to ensure that the average value of the spacing between clock pulses is equal to the desired sampling period, T_S, the instantaneous spacing between samples may vary greatly and unpredictably. This uncertainty in sample timing is termed *aperture jitter*.

In a communication system, aperture jitter causes uncertainty in the phase of the sampled signal, a rise in the data converter's noise floor, and an increase in the possibility of inter-symbol interference (ISI). These effects are directly proportional to the instantaneous rate of change in the input's signal voltage level, i.e., the signal's *slew rate*. Consequentially, higher frequency signals suffer more extensive signal quality reduction from aperture jitter than lower frequency signals as shown in Figure 4.8. Note that the aperture error for the high-frequency signal, Δv_{high}, is greater than the aperture error for the low-frequency signal, Δv_{low}, although the aperture jitter, τ_a, is the same for both signals.

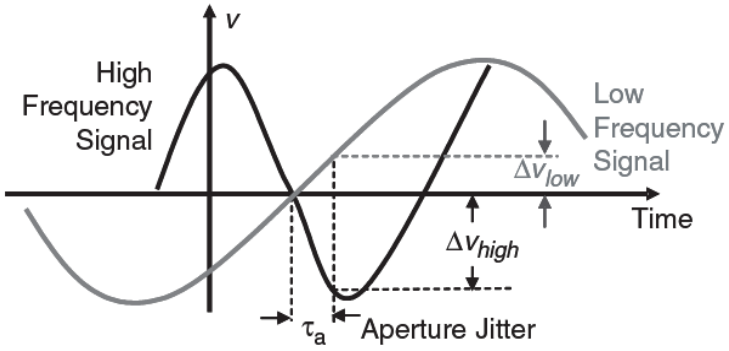


Figure 4.8: The measurement uncertainty due to aperture jitter varies with the signal's slew rate.

[Vasseaux_99] offers the following equation, based on the work of [Shinagawa_90], to describe the SNR of an aperture jitter limited system, i.e., a system in which quantization noise and other effects are negligible. In this case, the SNR of the system is given by

$$SNR = -10\log_{10}\left(2\pi^2 f^2 \overline{\Delta\tau_a^2}\right) \text{ (dB)} \qquad (4.12)$$

where f is the input signal's frequency and $\overline{\Delta\tau_a^2}$ is the expected variance in the aperture jitter. The effects of aperture jitter and input frequency are shown in Figure 4.9.

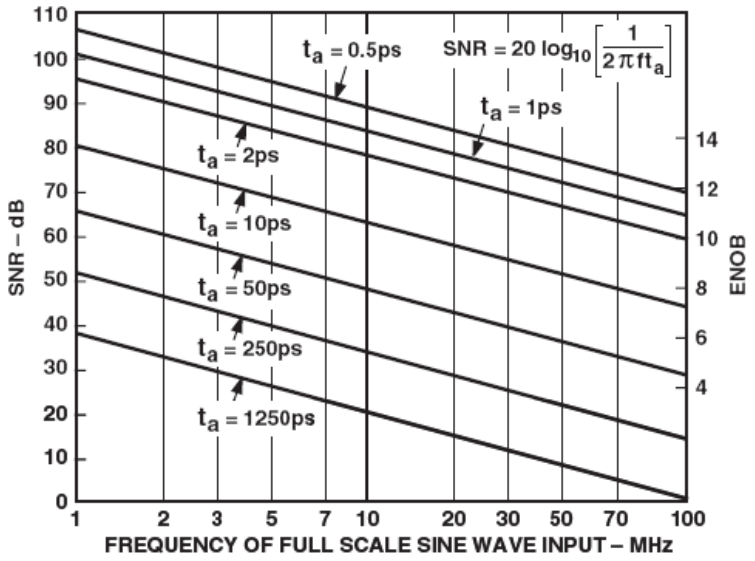


Figure 4.9: SNR Degradation Due to Aperture Jitter. [Brannon_00]

Note that as sampling and input bandwidths increase, aperture jitter becomes a limiting factor to the performance of an ADC. Also note that when analog bandwidth (see Section 4.1.1.5) is not a constraint, aperture jitter is typically the limiting factor to using undersampling due to the SNR degradation. As a rule of thumb, (0.12) can be added to (0.10) (SINAD) to calculate the effective dynamic range.

4.1.1.5 Analog Bandwidth

The RC circuits in the sample-and-hold circuitry cause a data converter to act as a lowpass filter, attenuating higher frequency input signals. The *analog bandwidth* of a data converter is the range of frequencies from DC to the point at which the spectral output of the fundamental swept frequency is reduced by 3 dB. Like SFDR, the analog bandwidth of a data converter varies with input signal power. Typically both small signal and FS signal measurements of analog bandwidth will be given for a data converter.

The analog bandwidth of a data converter does not guarantee good distortion metrics over the entire bandwidth as the SINAD ratio and thus the ENOB are also frequency sensitive and tend to degrade as the input frequency is increased [Mercer_01]. This effect is illustrated in Figure 4.10 where, although the full power bandwidth (FPBW) occurs at 1

MHz, the ENOB for an FS input experiences a bandwidth of around 100 kHz due to increased harmonic power. The ENOB for a signal backed off by 20 dB (≈3.32 bits), however, achieves the FPBW of the data converter.

In practice, the analog bandwidth of an ADC limits the highest input frequencies that can be safely recovered from an ADC. While most ADCs are designed such that the analog bandwidth is approximately the Nyquist rate, some ADCs have extended analog bandwidths (by adjusting the input capacitance) to permit undersampling.

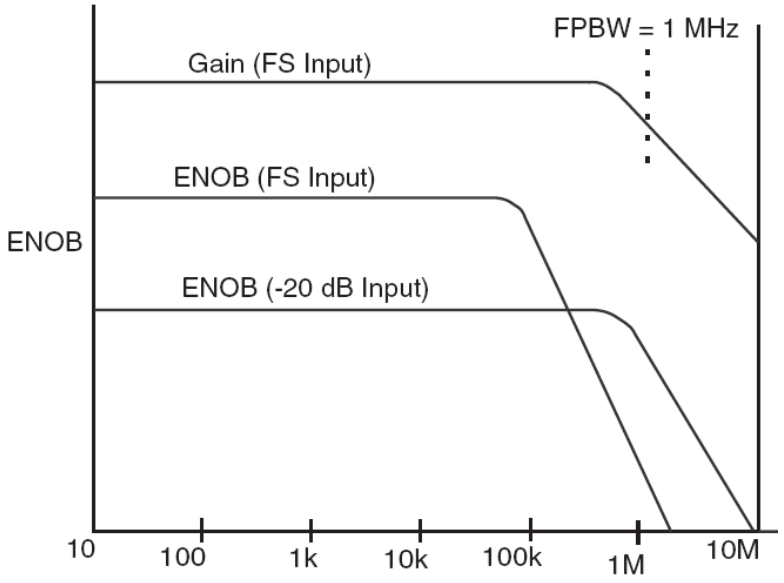


Figure 4.10: ENOB as a Function of Input Signal Level and Frequency. [Kester_96]

4.1.1.6 Power Consumption

[Kenington_00] derives an expression for power consumption based on the following four assumptions:

1. the quantizer itself does not consume any power,
2. a sample and hold circuit is used in the data converter,
3. the input signal is supplying the power to charge the sample and hold capacitance
4. the ADC's quantization noise power is equal to its thermal noise power.[2]

Assumption 1 is clearly false in any practical implementation (though useful to the applying the derivation to widely varying architectures), so this analysis forms a lower bound on device power consumption. From these assumptions, [Kenington_00] gives the

[2] This same condition is used to define the noise limit to ADC performance in Section 4.3.1.1.

following result for the minimum power consumption of an ADC with a sinusoidal input as a function of the sampling rate, effective temperature, and number of quantization bits.

$$Power > F_s kT_e 10^{(6B+1.76)/10} \; W \tag{4.13}$$

Note that a more general expression than is given in (0.13) can be found by substituting in a different PAPR (η) for

$$Power > F_s kT_e 10^{(6.02B+4.77-10\log_{10}\eta)/10} \; W \tag{4.14}$$

Note that while minimum power consumption is nominally independent of the free scale voltage, as the noise floor rises the full-scale voltage level is implicitly rising to preserve the effective number of quantization bits.

The relationship described in (0.13) can be visualized as shown in Figure 4.11 and Figure 4.12 for varying sampling rates and number of quantization bits for a $T_e = 288\ K$. According to the graph, an ADC with a twenty-bit resolution will theoretically consume a minimum of 600 mW when operating at 100 Msps and 6 W at 1 Gsps.

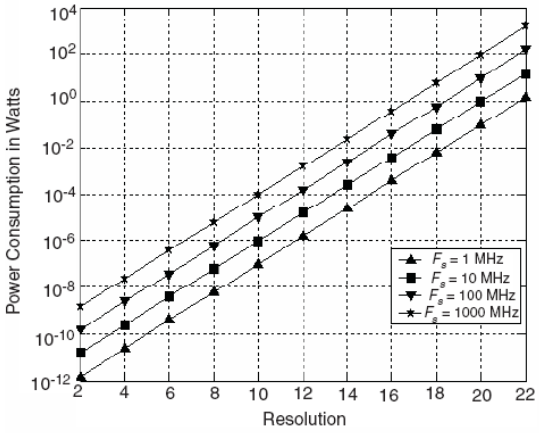


Figure 4.11: Power consumption lower bound (varying resolution)

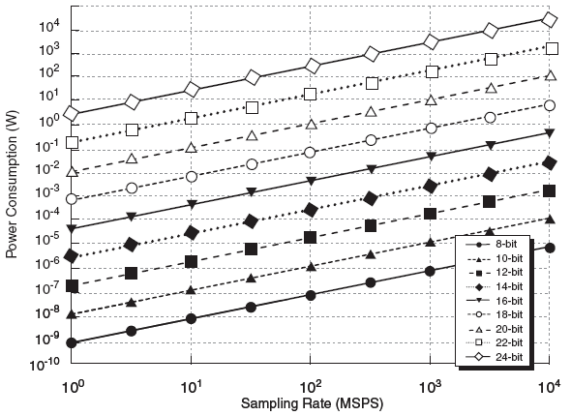


Figure 4.12: Power consumption lower bound (varying sampling rate) [Kenington_00]

This same linear relationship between sampling rate and power is borne out when examining how varying the sampling rate of ADCs within the same family (reclocking the same ADC) and examining the impact on power consumption. Note that while the plots are generally linear, there is significant variation in the slope of the power versus sampling rate curves.[3]

[3] Further information on the sources and methods used to collect this data is presented in Section 4.2.2.

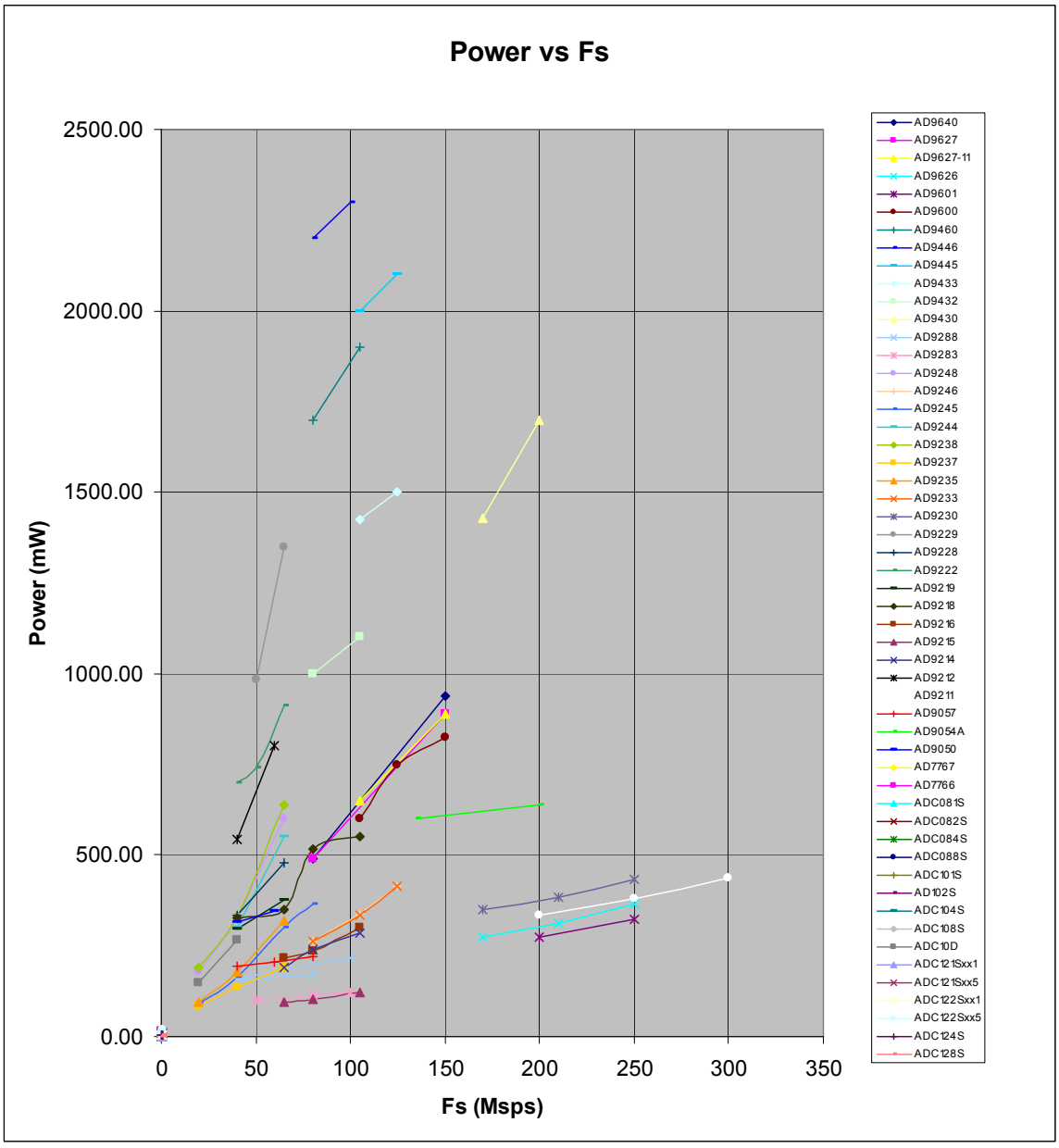


Figure 4.13: Relationship of Power and Sampling Rate within ADC Families

4.1.2 Relationship with Traditional SDR Performance Requirements

There exist well-defined relationships between ADC characteristics and traditional SDR performance requirements. The following discuss the relationships between sampling rates and signal bandwidth, quantization noise power and minimum detectable signal level, dynamic range and the operating environment, and power consumption.

4.1.2.1 Sampling Rate and Signal Bandwidth

For instance, given an input signal of bandwidth B_S, the minimum required sampling rate as given by Nyquist's sampling theorem is

$$F_S \geq 2B_S \tag{4.15}$$

Similarly when simultaneously recovering n signals, the following must hold.

$$F_S \geq 2\sum_{i=1}^{n} B_{S_i} \tag{4.16}$$

In practice, however, higher sampling rates are used to increase dynamic range and to aid the synchronization process. For an ADC with a fixed sampling rate (as most are), the minimum required sampling rate will be the largest sampling rate as specified by (0.16) [(0.15) is just (0.16) evaluated for n =1] for all the operating conditions the SDR is required to operate in.

4.1.2.2 Minimum Detectable Signal

The minimum detectable signal (MDS) is equivalent to the input channel noise power plus the cascaded noise figure. For an ADC, this is given by the sum of the quantization noise power N_Q as given in (0.7) and the thermal noise power, N_T. Note that some ADC architectures shape the quantization noise (e.g., sigma-delta ADCs) and thus would have a frequency-dependent MDS.

4.1.2.3 Dynamic Range

As illustrated in Section 4.1.1.3, dynamic range is required to recover a weak signal in the presence of a strong signal (or in the presence of many signals which aggregate to more undesired signal power). As wider RF chain bandwidths are used, more undesired signals will generally be input to the ADC.

In general, given n undesired signal and a signal which requires SNR_{min} (dB) for acceptable performance with the required dynamic range for the ADC is given by

$$DR = \max\left\{-10\log_{10}\left(\frac{R}{N_T + \sum_{i=1}^{n} I_i}\right), 0\right\} + SNR_{\min} \quad \text{(dB)} \tag{4.17}$$

where R is the specified minimum received power of the desired signal, N_T is thermal noise power, and I_i is the power of an undesired signal. Additional dynamic range can be required if it is desired to back-off the input power to the ADC to limit the probability of clipping by providing a safety margin for the total power of the input signal. In general, what constitutes an acceptable margin is a function of the η for the sum of all signals input to the ADC. Thus for proper SDR operation, an ADC with a SINAD (or the

equivalent ENOB) at least equal to (0.17) should be selected (plus any additional safety margin to minimize the probability of clipping).

4.1.2.4 Power Consumption

The power consumption of an ADC is additive with the power consumption of all devices in the SDR. A lower, device independent, bound for ADC power consumption was given by (0.14) and expressed in terms of the sampling rate, number of quantization bits, PAPR, and effective noise temperature.

4.2 Trends in ADC Performance

To analyze trends in ADC performance, an extensive Excel database was constructed. This was built on a previous database constructed by Virginia Tech students Bin Le and Thomas Rondeau which covered data converters up to 2004. The original database contained 914 parts and the updated database contains 1523 parts. The new database removes the price entries (deemed unreliable for purposes of comparing initial price points when examining current prices at a single point of time) and adds analog bandwidth entries for all parts (including previously surveyed parts).

The following manufacturers were surveyed: Analog Devices, Intersil, Linear, Maxim, MicroNet, Microchip, National Semiconductor, and Texas Instruments. Data in the following categories were collected for each ADC part: part #, architecture, nominal bits, sample rate, typical power consumption, bandwidth, SFDR, SINAD, ENOB, Substrate, # Channels, Normalized Power, and first year of manufacture.

Two primary types of data sources were used in constructing this database. The earlier 2004 database and datasheets sources from the identified manufacturers. The needed data sheets were downloaded from the ADC manufacturer websites and reviewed for relevant information.

4.2.1 Forces Driving ADC Trends

Commercially available parts are largely driven by the needs of the commercial market and improvements in fabrication technologies. While the latter is primarily driven by the needs of the processor industry, these improvements can usually be applied to manufacture of ADCs to allow smaller, faster, and in theory, lower power devices.

Because of the inherent tradeoffs in sampling rate, power consumption, and quantization bits, the performance metrics where these improvements are realized depends on the immediate needs of the market. For instance when GSM was being developed, there was a sharp upswing in the maximum number of quantization bits an ADC could support

because of GSM's large dynamic range requirements at the expense of sampling rates. As commercial signals have started trending towards wider bandwidth and multi-band signals, commercial ADCs have increased their sampling rates and increased the analog bandwidths at the expense of the number of quantization bits.

4.2.2 Review of Key ADC Trends

Using the database described in the preceding, we analyzed trends in sampling rates, quantization bits, bandwidth, power consumption, and performance.

4.2.2.1 Sampling Rate Trends

Using our database of commercially available ADCs, we generated the scatter-plot of sampling rates versus year of initial manufacture shown in Figure 4.14 where each ADC architecture has been encoded with a different symbol.

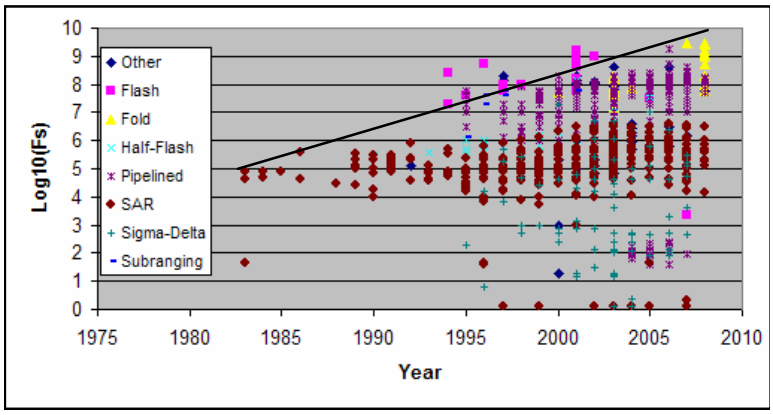


Figure 4.14: Maximum Sampling Rates Have doubled at a rate of once every 1.5 years when referenced to 1981, but at a rate of once every 3.5 years when referenced to 1994.

Impact of architecture

The architecture strongly influences conversion speed with the recursive (SAR, sigma-delta) and multi-stage (e.g., pipelined) architectures generally running slower than the single-stage architectures (e.g., Flash).

Peak Speed

Depending on the chosen reference period, peak ADC rates have either increased tremendously or at a rate commensurate with the gains seen in transistors. For instance, in 1981, the fastest surveyed ADC had a clock rate of 25 kHz, while in 2007, the fastest

surveyed ADC had a clock rate of 3 GHz.[4] This implies approximately 17 doublings in clock rate in 26 years or approximately every year and a half. However, in 1994, the fastest clock rate was 250 MHz which gives about 3.5 doublings in 13 years or equivalently, **peak F_S approximately doubles every 3.5 years**.

As this latter rate is more inline with the progression of transistor technologies where chip speed doubles every 3 years, it may be the case that the limited data sets for early 80s ADCs are not extensive enough to be representative, a not unsurprising implication given the methodology of consulting online databases of ADC datasheets.

4.2.2.2 ENOB Trends

As shown in Figure 4.15 there has been noticeably less progress in the effective number of bits, though there is significant variation between architectures and the appearance of sigma-delta converters coincides with a significant increase in the peak ENOB. The lack of significant growth in bit rates is particularly noticeable when considering the best linear fit to the data for each architecture as shown in Table 4.1. Note that since ENOB is derived from an ADC's SINAD, which is generally slightly lower than an ADC's SFDR, this is also loosely approximating trends in SFDR. While changes in ADC architectures led to increased ENOB, **ENOB trends within architectures are flat**. This is likely due to existing number of quantization bits being sufficient for existing applications.

Table 4.1: Slopes of best linear fits for ENOB for varying architectures.

Architecture	Δ ENOB / yr
Flash	0.025892747
Fold	0.028201767
Half-Flash	-0.193450143
Pipeline	-0.000168524
SAR	0.063200717
Sigma-Delta	0.063024667

[4] Note, these need not be the fastest ADCs available either today or in 1981. In fact, some others are known to the authors. However, for the survey methods used (reviewing datasheets for ADCs available online from major commercial ADC vendors), this is the extent of the capabilities.

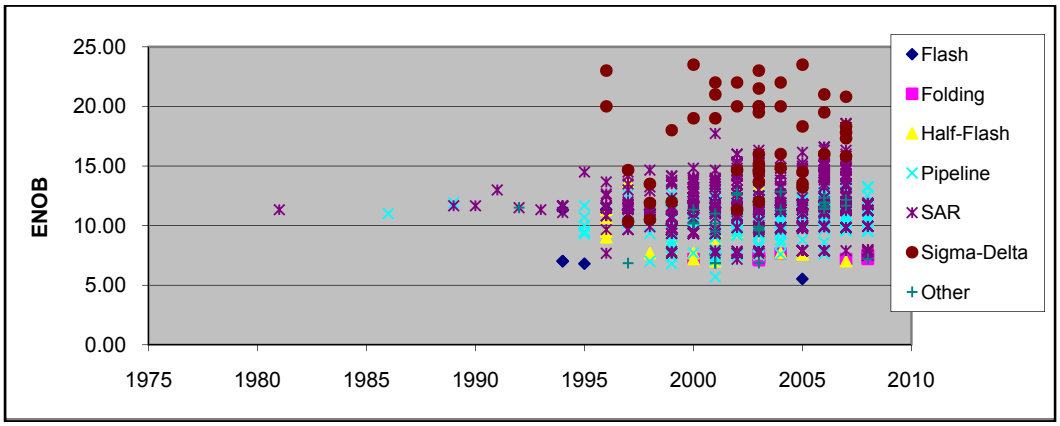


Figure 4.15: Most of the growth in quantization bits has come from shifts in architectures rather than improvements in fabrication technology.

4.2.2.3 Analog Bandwidth Trends

Using our database of commercially available ADCs, we generated the scatter-plot of full-power analog bandwidths versus year of initial manufacture shown in Figure 4.14 where each ADC architecture has been encoded with a different symbol. Note that our data base did not have data sheets that reported an analog bandwidth prior to 1992. While a wide range of bandwidths are still encountered, there is a general trend upwards across in the data.

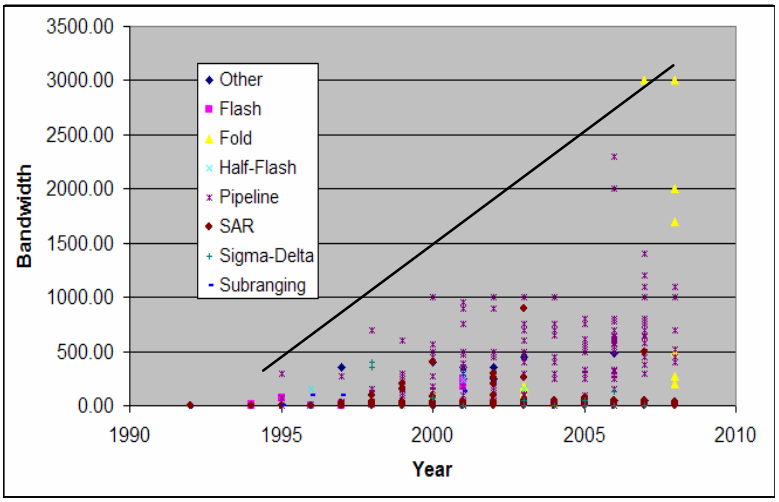


Figure 4.16: Analog Bandwidths

In 2007, the widest full-power bandwidth was 3 GHz; in 1992 this was 1 MHz. Technically, this corresponds to a doubling rate of every 1.3 years, but this should not be a direct result of leveraging the fabrication technology trends as in theory any effective input bandwidth could be designed.

Finally, note that the fastest data converters currently have full-power bandwidths equal to their sampling rates. This means that in current practice, it is impossible to perform bandpass sampling in the fastest data converters more than 1 Nyquist zone above baseband. However, a 3 GHz bandwidth does capture most of the current commercial wireless standards which tend to operate below 2.5 GHz.

4.2.2.4 Power Consumption Trends

A scatter plot of typical reported power consumption measurements is shown in Figure 4.17 where there is a noticeable upwards trend in the average and maximum power levels. Performing a line fit for each architecture (a reasonable assumption for growing sampling rates and near constant ENOB) yields the results shown in Table 4.2. This trend is much more marked when considering the maximum power consumed by each ADC where in 1989, 135 mW was consumed with 7 W in 2003 (490 mW / yr).

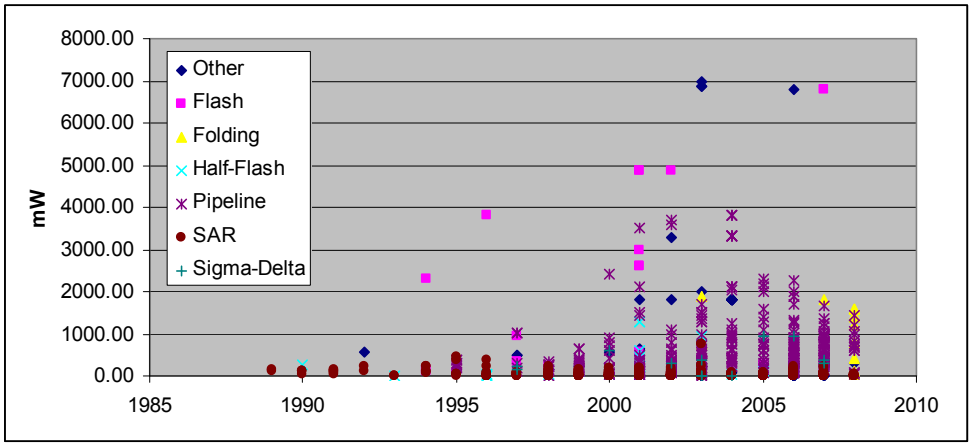


Figure 4.17: ADC Power Consumption has been Increasing over Time

Table 4.2: Slopes of best linear fits for Power Consumption for varying architectures

Architecture	Δ mW / yr
Sigma-Delta	38.46136
SAR	-3.38049
Pipeline	36.76544
Half-Flash	18.14782
Folding	152.6672
Flash	312.0247
Other	16.89148

4.2.2.5 Composite Performance Trends

Because of the tradeoff between quantization levels and sampling rate illustrated in Figure 4.18, a commonly used metric for ADC performance is the product of the effective number of quantization levels (2^{ENOB}) and sampling rate (F_S) and is expressed as

$$P' = 2^{ENOB} F_S \quad (4.18)$$

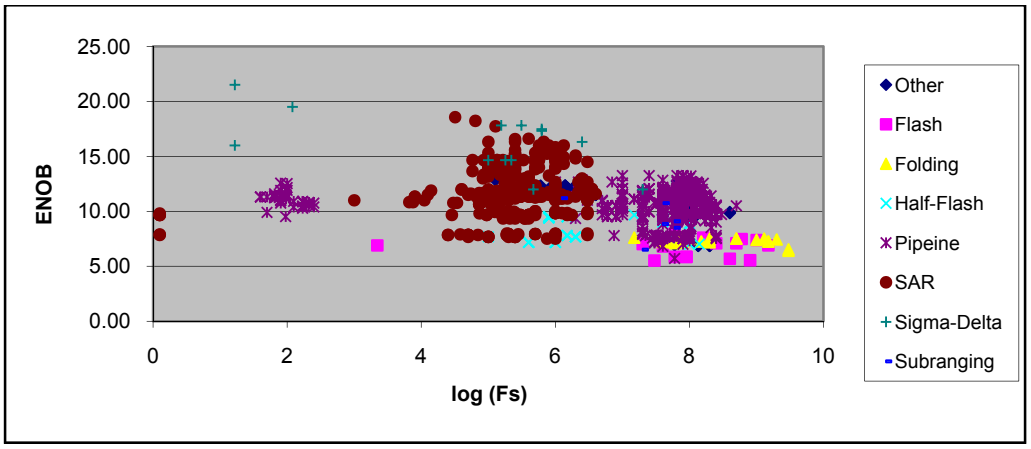


Figure 4.18: There exists a fundamental tradeoff between sampling rate and number of quantization bits.

Over time, there has been a general increase in *P* as shown in Figure 4.19. For example, in 1996, the peak *P'* increased by a factor of 15 from 1996 to 2008. **Thus peak ADC performance has been improving by about 25% / year**. As noted in Sections 4.2.2.1 and 4.2.2.2, this gain is largely due to improvements in the sampling rate, which implies the gains are attributable to improvements in fabrication technologies.

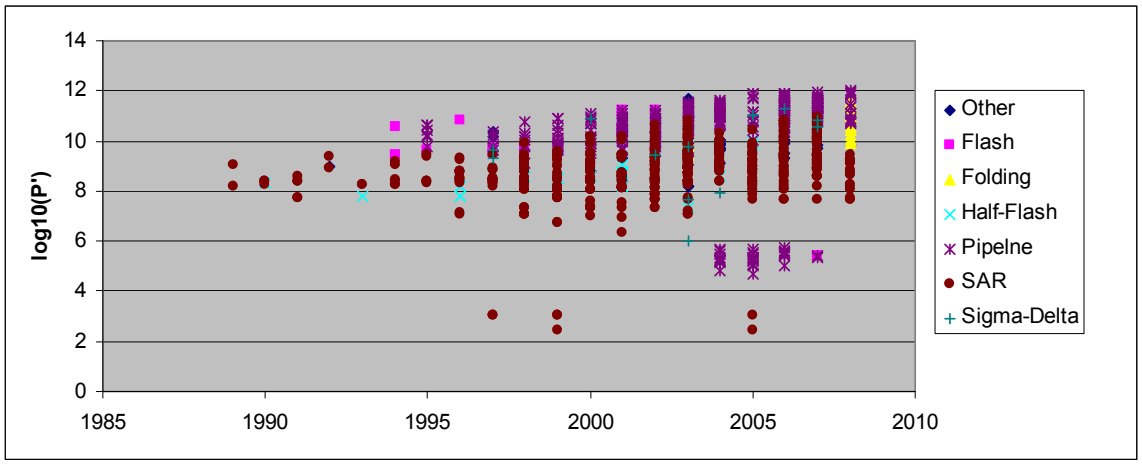


Figure 4.19: *P* has increased significantly over time.

However, this increase in performance has been accompanied by increases in power consumption. Recognizing this fact, another commonly used metric is defined as the ratio of *P'* to power consumption or

$$F = 2^{ENOB} F_S / Power \tag{4.19}$$

F has also seen significant improvements in performance as shown in Figure 4.20. As ENOB has been largely flat, this implies that ***F_S* is growing faster than power consumption**. This is significantly different than the trend seen for General Purpose Processors (GPPs) where increases in clock rates have been accompanied by increases in power. This is likely due to ADC vendors making different choices in scaling factors (feature size and transistor voltages) motivated by the need for lower power handsets.

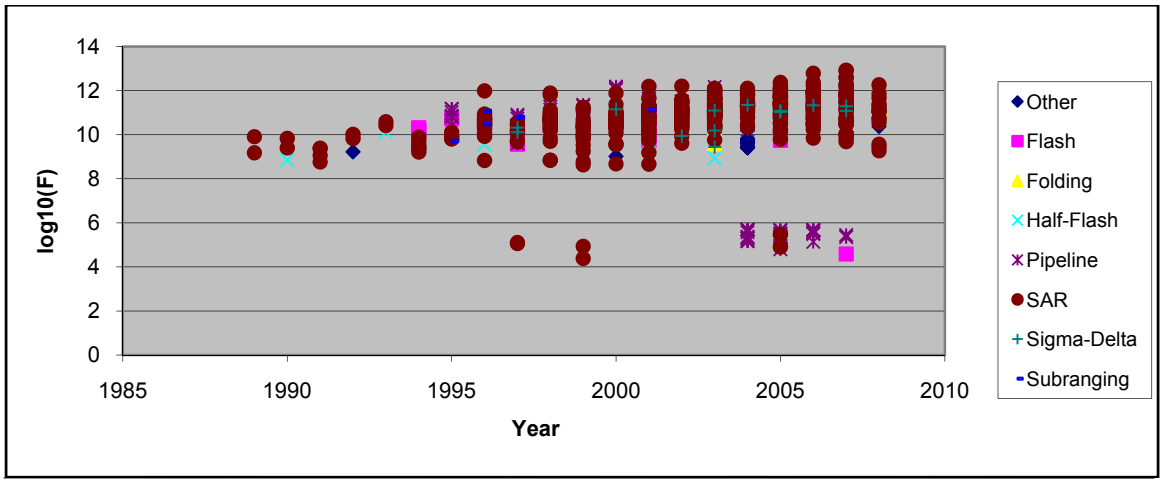


Figure 4.20: Even when considering power consumption, ADCs have seen significant improvements in performance.

4.2.3 Implications for SDR Design and Implementation

The preceding noted that over a long period sampling rates have been doubling at a rate faster than once every two years. More recently, however, this rate of increase has slowed to be more in-line with the speed improvements seen with transistors (doubling approximately every 3 years). Power consumption has also been increasing, but at a slower rate than sampling rates so that overall ADC efficiency has been improving even though ENOB has been relatively flat.

Overall, peak ADC performance has been increasing at a rate of about 25% per year. With current commercial ADC sampling rates at 3 GHz, it appears that sampling rates in and of themselves will not be a significant impediment to implementing wide-band SDR. However as we will show in Section 4.3, fabrication and physical limits will mean that an

ADC that both digitizes a wide bandwidth and exhibits a high dynamic range will be impossible.

4.3 Fundamental Limits to ADC Performance

The following discusses the fabrication and physical limits to ADC performance and the implications to SDR design and implementation.

4.3.1 Sources of Fundamental Limits

As shown in Figure 4.18, there is a fundamental tradeoff between ENOB and F_S. In effect, ADC performance, *P*', is allocated between ENOB and F_S as a result of the specific ADC architectures and the implementation choices made for those architectures. For example, a SAR ADC is explicitly exchanging clock rate to increase resolution, and fabrications limits, e.g., while a single-stage Flash maximizes conversion rate, there are limits to how accurately the voltage can be subdivided by a voltage ladder.

However, there are fundamental limits to how large *P'* can be based on effects such as aperture jitter, input noise, and comparator ambiguity. The following discusses the role these limits play in the total performance as well as between achievable sampling rates and dynamic range (ENOB).

4.3.1.1 Performance Limits

[Walden_99] highlights three key fabrication factors that limit ADC performance: input noise effects, timing jitter, and comparator ambiguity. Input noise is generally dominated by thermal noise which is caused by the random motion of electrons and aperture noise is the result of randomness in the sampling process. Comparator ambiguity is the result of measurement errors in the quantization process due to physical limitations on the regeneration time of transistors in the comparators. Assuming that each factor is independent and is the only limiting factor, [Walden_99] derives the following equations to ascertain when an uncertainty of ±0.5 LSB is induced thereby limiting ADC performance.

When only considering noise input to the ADC (thermal, shot, flicker) maximum ENOB is given by

$$ENOB_{Thermal} = \log_2\left(\frac{V_{FS}^2}{6kTR_{eff}F_S}\right)^{1/2} - 1 \qquad (4.20)$$

where *T* is temperature in Kelvin and R_{eff} captures the effect of all external noise sources. When aperture noise (jitter) is the only limiting factor, maximum ENOB is given by

$$ENOB_{jitter} = \log_2\left(\frac{2}{\sqrt{3}\pi F_S \tau_a}\right) - 1 \tag{4.21}$$

When only comparator ambiguity (uncertainty if a signal is above or below a desired level) is a consideration, maximum ENOB is given by

$$ENOB_{comp} = \frac{\pi f_T}{6.93 F_S} - 1.1 \tag{4.22}$$

where f_T is the regeneration rate of the transistors used in the ADC's comparators.

Another key result of Walden's equations is when each of the factors dominates performance.

The key parameters of each of these factors – effective resistance R_{eff}, aperture jitter τ_a, and regeneration rate f_T – depend on the state of the art in fabrication technology. However, these limits are arguably not fundamental because as fabrication technology slowly improves these limits will slowly relax.

To define a physical bound to ADC performance, consider the energy-time formulation of the Heisenberg uncertainty principle shown in (0.23)

$$\Delta E \Delta t > h / 2\pi \tag{4.23}$$

where ΔE is the uncertainty in the measurement of a system's energy, Δt is the time uncertainty in the measurement (loosely, the period over which measurements are not made), and h is Planck's constant (6.6217×10^{-34} Joules · seconds). For an ADC, Δt is half the sampling time, and ΔE is the power uncertainty equal to half a quantization level integrated over Δt. Substituting these expressions into (0.23) yields

$$\Delta E \Delta t = \frac{\left(V_{FS} 2^{-B} / 2\right)^2}{R} \frac{T_s}{2} \frac{T_s}{2} = \frac{\left(\frac{V_{FS} 2^{-B}}{4 f_s}\right)^2}{R} > h / 2\pi \tag{4.24}$$

where R is the ADC impedance. In Walden's notation, this can be expressed as a limit on the number of quantization bits as shown in (0.25) where it is implicitly assumed that $SQNR\ (2^{-B}) = SINAD$ which means that in practice, ENOB will be even smaller.

$$ENOB < \log_2\left(\frac{V_{FS} / 4}{F_s \sqrt{R \frac{h}{2\pi}}}\right) \tag{4.25}$$

If we assume an R of 50 Ω and a V_{FS} of 2.5 V, then P' can never exceed 8.61×10^{15}. This corresponds to a 5 GHz ADC with 20.7 effective bits or 14 effective bits with a

maximum sampling rate of 526 GHz. As the best commercially available ADC has a *P'* of 9.9640×10^{11}, at the current improvement rate of 25% ADC performance could improve for the next 40 years ($\log_{1.25}(8.61 \times 10^{15} / 9.9640 \times 10^{11})$) before reaching the Heisenberg uncertainty principle. However, we are only projecting another 16 years of improvements in transistor technologies, which means that **ADC performance improvements should continue for the next 16 years**. This would result in a *P'* of about 3.54×10^{13} – two orders of magnitude less than the estimated Heisenberg limit.

To visualize where these fabrication and physical limits fall in the ADC trade space, we need to assign values to the state-of-the art in timing jitter (τ_a) and comparator regeneration rate (f_T). Of the surveyed ADCs, the smallest value for τ_a was 60 fs in the AD9445 [AD9445_05]. Note that the processor with the fastest sampling rate, the ADC083000 [ADC083000_07] with a sampling rate of 3 GHz, had a value for τ_a of 550 fs. This induced to a significant degradation in SINAD at higher input frequencies so that near Nyquist the rate (1498 MHz), a SINAD of 41.1 dB was reported while at 373 MHz, a SINAD of 45 dB was reported. Also of note, the AD9445 was first manufactured in 2005, so significant improvements in aperture jitter should be possible.

Unfortunately, ADC datasheets do not generally report comparator regeneration times. So we initially consulted Analog Devices' most recent comparator selection guide [AD_07] for stand-alone comparators where the minimum pulse width was 80 ps for the ADCMP582 for a rate of $f_T = 12.5$ GHz. Unfortunately, this value did not bound the comparator regeneration time as several converters fell outside of this bound. Since a comparator's regeneration rate would be bound by transistor switching rates, another approximation for comparator regeneration time is the oscillation time for a ring-oscillator (an odd number of inverters arranged so that the output of the last inverter is the input to the first inverter). [Mistry_07] reported that in Intel's 45nm process operating under typical conditions a per-stage delay of 5.1 ps was seen ($f_T = 196$ GHz) in a ring oscillator. We then used 5.1 ps for our estimate of the current limit to comparator regeneration time.

Substituting these numbers along with an effective resistance of $R_{eff} = 50\ \Omega$, and a temperature $T = 298$ K, and $V_{FS} = 2.5$ V into (0.20), (0.21), (0.22), and (0.25) and plotting these curves on the ENOB vs F_S scatter plot from the ADC survey yields Figure 4.21.

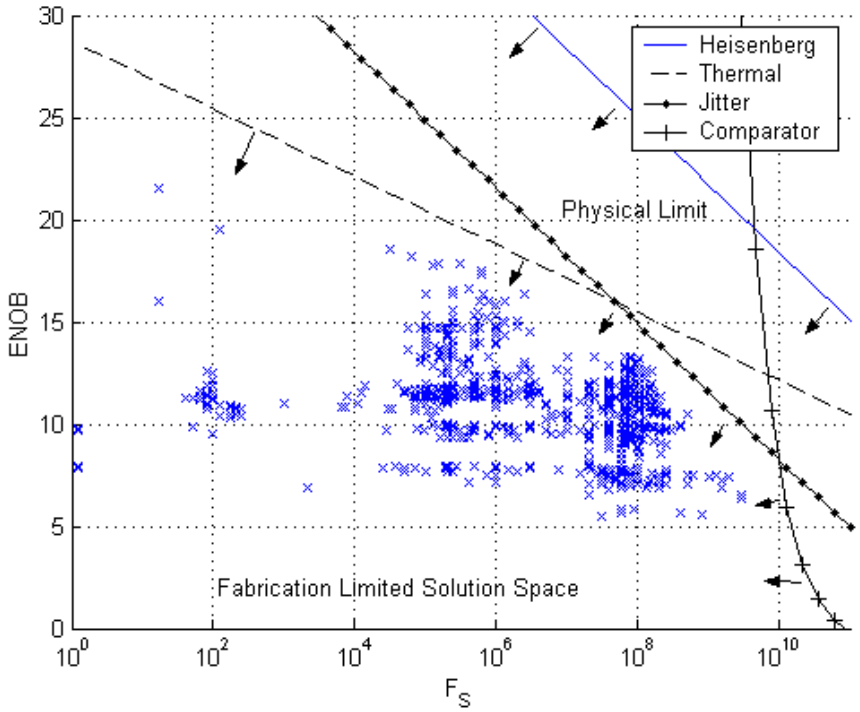


Figure 4.21: The ADC solution space is constrained by fabrication limits and physical limits.

As expected, all surveyed ADCs fall within the bounds predicted by the estimated limits to fabrication technologies with different sampling rate regimes dominated by different fabrication limits. The approximate boundaries for these fabrication limits are listed in Table 4.3. Note that most current SDR designs fall in the sampling rate region where performance is limited by aperture jitter, so for the purposes of SDR, **aperture jitter is the primary limiting factor to ADC performance in SDR applications**.

Table 4.3: Limiting Factors by Sampling Rate Region

F_S region	Fabrication Limit
0 to 44 MHz	Input noise (thermal, shot)
44 MHz – 10 GHz	Aperture Jitter
10 GHz - ∞	Comparator regeneration

4.3.1.2 Power Limits

By substituting the relationship $Power = V_{FS}^2 / R$ into (0.25), we can derive an expressions relating ADC performance and power consumption as constrained by the Heisenberg uncertainty principle as shown in the following.

$$P' = F_s 2^{ENOB} < \frac{\sqrt{Power}}{4\sqrt{h/2\pi}} \tag{4.26}$$

$$(16h/2\pi)P'^2 < Power \tag{4.27}$$

This implies that near the Heisenberg limit, if ADC performance improves by a factor *n*, **ADC power consumption will increase on the order n^2**. This trend is reflected in current ADC trends. Figure 4.22 plots (0.27) on a scatter-plot of surveyed power consumption versus device performance ($F_S\, 2^{ENOB}$).

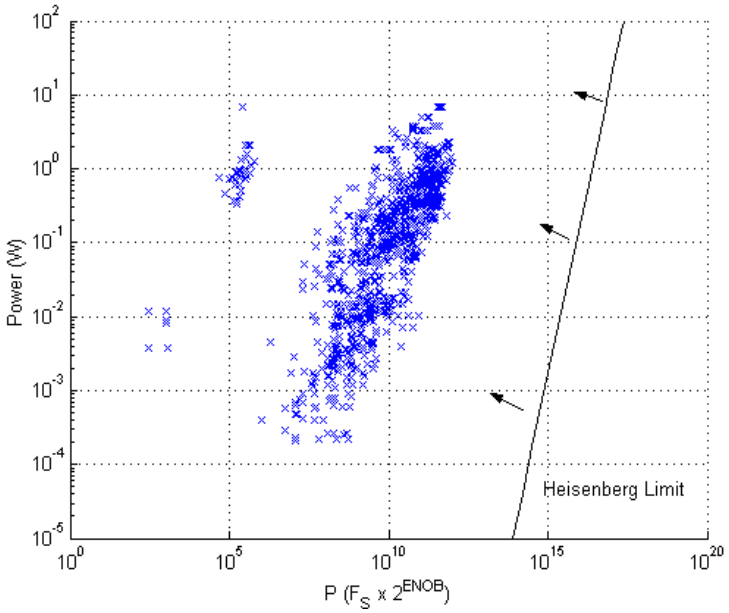


Figure 4.22: Physics limits minimum power consumption as a function of ADC performance.

4.3.2 Implications for SDR Design and Implementation

In general, there is a significant variation between waveforms to the required digitization bandwidth and required ENOB. Thus different architectures and different ADCs are frequently employed to handle different waveforms. For an SDR expected to implement many different waveforms, employing an ADC architecture that allows reconfiguration to dynamically emphasize either ENOB or F_S (e.g., SAR or pipelined) should be valuable.

Frequently, the ideal SDR architecture is envisioned as an ADC operating next to the antenna. As a quick assessment of the feasibility of such an architecture, consider recovering a GSM signal in the GSM-900 band which at a minimum requires for digitization of 200 kHz and a dynamic range of 88 dB. If we naively assume this 88 dB dynamic range requirement holds for signals digitized over the entire 25 MHz GSM-900 block so that we wish to recover a -101 dBm signal in the presence of a -13 dBm signal, and assume that every possible 25 MHz block digitized signal is dominated by a single

-13 dBm signal, then the total input power is -13 dBm + 10 $\log_{10}$ (N) for digitizing across N 25 MHz blocks.

So for N 25 MHz chunks, at least a sampling rate of 50xN MHz is required. The required dynamic range is then 88 dB + 10 $\log_{10}$ (N). Assuming a total of 12 dB is required for proper signal recovery and to account for AGC circuit variation, the required ENOB for this hypothetical ADC is:

$$ENOB \approx \left[100 + 10\log_{10}(N)\right] / 6.02 \tag{4.28}$$

The required ADC performance can be estimated as

$$P' = 2N \times 2^{\left[100+10\log_{10}(N)\right]/6.02} \tag{4.29}$$

This curve is plotted versus digitized bandwidth along with the Heisenberg limit for P' identified above (8.61×10^{15} for $R = 50\ \Omega$, V_{FS} =2.5 V), the best available P' (9.9640×10^{11}), and the projected best available P' when Moore's Law ends (16 years) Note that under these assumptions, even expanding the digitized bandwidth to 25 MHz (from 200 kHz) slightly exceeds the limits of the performance of the best currently available commercial ADC and at our projected realizable limit, only about 50 MHz will be possible. Further, attempting to digitize from DC to 2.5 GHz would violate the performance limit imposed by Heisenberg's uncertainty principle.

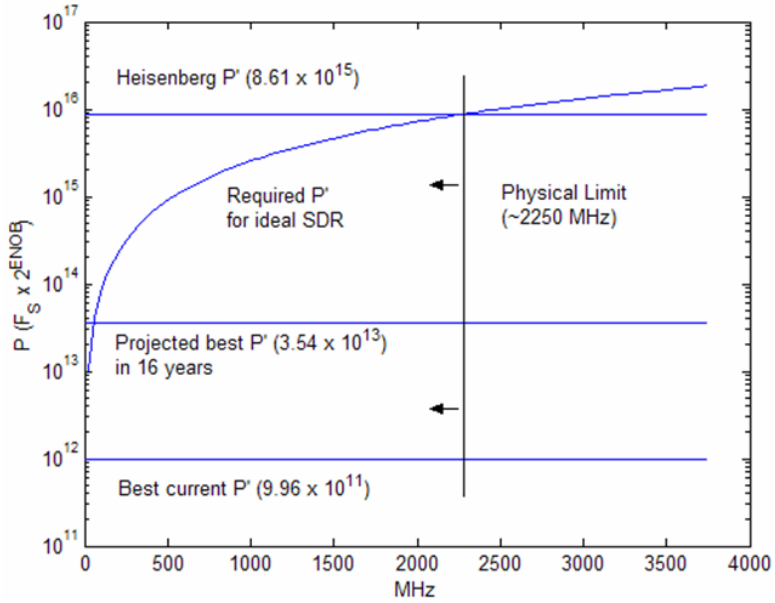


Figure 4.23: There is a physical limit to how much bandwidth can be digitized for a specified dynamic range.

For the following reasons, this brief analysis is likely underestimating the performance requirements for an ADC supporting the ideal SDR:

- As digitized bandwidth increases, the number of digitized signals increases and $\eta \rightarrow \infty$ [see equation (0.5)].
- The GSM dynamic range specification is for in-band cell-phones whose transmission power is constrained to significantly less than the power limits for other transmitted signals (e.g., base stations and broadcast towers).

Because of these reasons and based on this analysis, **it is our judgment that the ideal SDR (simultaneous digitization at the antenna across the band from DC to light) is not physically possible**. Further, as a first approximation, it appears that digitization at the antenna from DC to 2.5 GHz is also impossible and not an artifact of current fabrication limitations.

Thus, for SDR to achieve a wideband front-end, it will be necessary to employ a tunable RF front end that limits the performance requirements placed on the ADC.

REFERENCES

[AD_07] Analog Devices, "Comparator Selection Guide 2007," Available online: http://www.analog.com/UploadedFiles/Selection_Tables/595625874Comparator_Brochure_07.pdf
[AD9445_05] "AD9445 DataSheet Rev 0," Analog Devices, 2005.
[ADC083000_07] "ADC083000 DataSheet," National Semiconductor Corporation, May 2007.
[Brannon_96] B. Brannon, "Wideband Radios Need Wide Dynamic Range Converters," *Analog Dialogue*, vol. 29, no. 2, 1996.
[Brannon_00] B. Brannon, "Aperture Uncertainty and ADC System Performance," Tech. Rep. AN-501, 2000.
[Kenington_00] P. B. Kenington and L. Astier, "Power Consumption of A/D Converters for Software Radio Applications," *IEEE Transactions on Vehicular Technology*, vol. 49, March 2000.
[Kester_96] W. Kester, "High speed design techniques," tech. rep., Analog Devices, Inc., 1996.
[MAX_00] "ADC and DAC Glossary," tech. rep., MAXIM Integrated Products, December 2000.
[Mercer_01] C. Mercer, "Acquisition Systems, Number of Bits, and Dynamic Range," tech. rep., PROSIG Signal Processing Tutorials, June 2001.
[Mistry_07] K. Mistry, et al., "A 45nm Logic Technology with High-k + Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," 2007 International Electron Devices Meeting, International Electron Devices Meeting Technical Digest, paper 10.2, 2007.
[Mitola_95] Joseph Mitola, III, "The Software Radio Architecture," *IEEE Communications Magazine*, vol. 33, pp. 26–38, May 1995.
[Reed_02] J. Reed, J. Neel., and Sujayeendar Sachindar, "Analog-to-Digital and Digital-to-Analog Conversion," in J. Reed, *Software Radio: A Modern Approach to Radio Engineering*, Prentice Hall, 2002.
[SDRF] www.sdrforum.org.
[Shinagawa_90] M. Shinagawa, Y. Akazawa, and T. Wakimoto, "Jitter Analysis of High-Speed Sampling Systems," *IEEE J. of Solid-State Circuits*, vol. 25, February 1990.
[Telecom_96] "Telecommunications: Glossary of Telecommunications Terms," tech. rep., Institute for Telecommunication Sciences, 1996. Federal Standard 1037C.
[Vasseaux_99] T. Vasseaux, B. Huyart, P. Loumeau, and J. F. Naviner, "A Track & Hold Mixer for Direct Conversion by Subsampling," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 584–587, 1999.
[Walden_99] R. H. Walden, "Analog-to-Digital Converter Survey and Analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, April 1999.

5 Digital Signal Processors

Beginning in 1979 with the Air Force's Tactical Anti-Jam Programmable Signal Processor (TAJPSP) program, there has been a migration of radio architectures from those that implemented functionality with analog and integrated circuit components to architectures that use general purpose processors (GPP), digital signal processors (DSP) (micro-processors optimized for digital signal processing applications), and field programmable gate arrays (FPGA) to perform the bulk of the operations. The waveforms of radios designed using these components are not defined by hardware; rather, it is the software loaded onto the GPP, DSP or FPGA that determine the radios' waveforms.

The recognition of this paradigm shift – from hardware to software implementations – led to the coining of the term "software radio" by Joe Mitola in 1991 to differentiate radios that are implemented primarily in software from those implemented principally in hardware. More formally, the Software Defined Radio Forum defines a *software defined radio* [SDRF_08] as

> "A radio for which software processing is used to implement some or all of the physical layer processes."

Thus microprocessors (both GPPs and DSPs) are central to the concept of software radio and the capabilities of these processors largely control the percentage of radio operations that can be performed in software.

The choice of a processor is generally influenced by the following considerations:

- **Maximizing computational capacity** – Abstractly, each waveform an SDR supports requires a minimum number of operations to be completed per unit time. In general, the faster these computations are completed, the greater the number of waveforms the SDR can support. In practice, however, what constitutes a fundamental unit of computation varies significantly from processor to processor.
- **Minimizing power consumption** – In general, all SWAP (size-weight and power) considerations factor into the selection of every component in an SDR. However, with power consumption levels that can extend into the hundreds of Watts, processor power consumption is generally an important consideration in SDR design.
- **Minimizing reconfiguration time** – Software radio, and particularly cognitive radio applications, can be limited by systems that are slow to change operation. Prior to development of partially-reconfigurable FPGAs, this was a primary rationale for choosing DSPs (reconfiguration times on the order of ns) over FPGAs (reconfiguration times on the order of ms). However, as this study focuses

on DSPs and GPPs, for which reconfiguration time is not a significant concern, this issue is not explored in depth.

- **Software considerations** – Ease of development, code-reuse, code portability, and code maintainability are all influenced by the choice of processing architecture. However, these issues are beyond the scope of this study and will generally not constitute a fundamental limit to SDR deployment.

The primary focus of this study will be on computational capacity and power consumption of GPPs and DSPs. [5] In general, computational capacity is positively correlated with power consumption so that there is a fundamental tradeoff between maximizing computational capacity and minimizing power consumption as illustrated in Figure 5.1. Additionally, this relationship is strongly influenced by transistor fabrication technology and processor architecture such that any limit imposed by fabrication technology. To further complicate matters, the relative computational efficiency of a processor varies from waveform to waveform such that an efficient processor for implementing one waveform may be inefficient for implementing another.

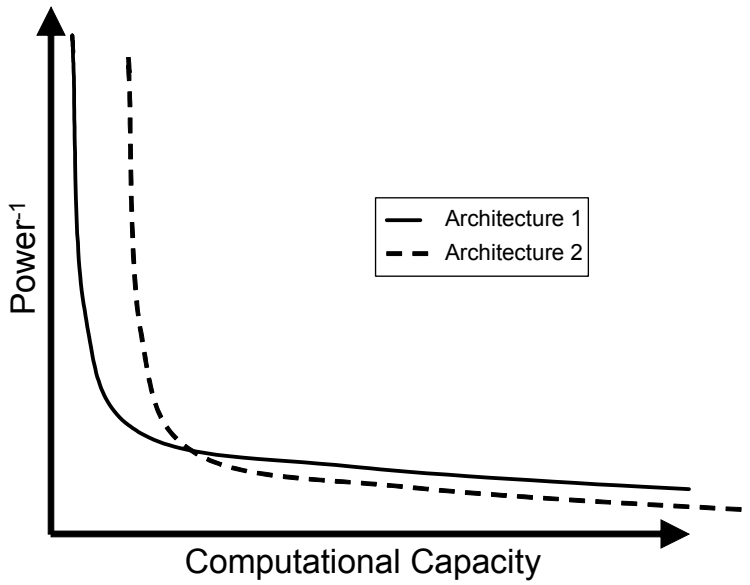


Figure 5.1: The fundamental tradeoff between maximizing computational capacity and minimizing power is influenced by processor architecture and the waveforms that are being implemented.

[5] FPGAs will be briefly addressed as relevant.

This study uncovered the following key findings.

- While the growth in GPP clock speeds have leveled-off, this has not yet happened for DSPs.
- The benefits of Moore's Law are being realized in increasing parallelism which is dramatically improving computational efficiency.
- The picoChip products are significantly ahead of traditional DSPs in terms of computational efficiency.
- Cost may become a significant limiting factor in the near future as tool costs continue to rise exponentially while revenues have flattened.
- At current rates, Moore's Law faces a physical limit in 16 years.
- Based on our transistor projections and a hypothetical "bare-bones" DSP, we estimate a limit of 15 nW per MMACS.

Figure 5.2: Summary of key findings by this study.

The remainder of this Section discusses how we came to these conclusions with the following structure. Section 5.1 reviews the relationships between key DSP characteristics and their relationship to SDR performance. Section 5.2 reviews the forces driving DSP design and trends in key DSP and transistor performance metrics. Section 5.3 estimates fundamental limits to transistor and DSP performance and discusses how these limits will impact SDR design and implementation.

5.1 Relationships of Key DSP Characteristics

The following reviews key DSP characteristics, their relationships with each other and their relationships with typical SDR performance metrics.

5.1.1 Overview of Key DSP Characteristics

The most critical aspects to the performance of a DSP are its power consumption and its computational capacity. As conceptually illustrated in Figure 5.1, power consumption is positively correlated with the computational capacity of a processor. So all else being equal, these two opposing objectives imply that an SDR designer should choose the processor that can satisfy the complexity requirements of the most complex waveform intended for deployment on the SDR with the least power. Additionally, this relationship is strongly influenced by transistor fabrication technology and processor architecture.

5.1.1.1 Computational Capacity

Computational capacity refers to the maximum number of computations a processor can support per unit time. At first glance, this appears to be largely determined by the clock

rate of a processor and indeed this was how GPPs were marketed for years. However, different processors will execute a different number of instructions per cycle. To reflect this fact, another commonly metric is **M**illions of **I**nstructions **P**er **S**econd (**MIPS**), which is also sometimes used to describe the complexity of waveforms. For example [Crocket_98] provides the estimates for computational complexity for WCDMA, IS-95, IS-136, and GSM as shown in Table 5.1.

Table 5.1: Estimated Computational Complexities for Selected Waveforms [Crocket_98]

Waveform	Approximate Complexity
WCDMA	5000 MIPS
IS-95	500 MIPS
IS-136	200 MIPS
GSM	100 MIPS

However, applying these metrics to processors can lead to misleading results as what constitutes an instruction or an operation can vary significantly from processor to processor. For instance, a processor that includes dedicated circuitry for the Viterbi and Turbo Decoder operations in WCDMA (e.g., the TMS3206416T from Texas Instruments) could call an entire decoding process with a single instruction while another processor would require many more instructions to achieve the same result. Then even with a higher MIPS rating, the second processor may not be able to implement the WCDMA waveform while the first would have no problem implementing WCDMA.

A related expression to MIPS that seemingly overcomes this difficulty is the number of operations per second, e.g., **M**illions of **O**perations **P**er **S**econd (**MOPS**), and **M**illions of **FL**oating-point **O**perations **P**er **S**econd (**MFLOPS**). Independent of how the instructions are issued to implement an algorithm, complex processes can be expressed in terms of a number of different kinds of operations. For instance, Table 5.2 gives parameterized estimations of the number of operations required to implement a number of different common waveform components. A minimum operation rate can then be calculated by dividing the number of operations by the required time for each process. In theory, this could then be compared against the product of the number of operations a DSP performs per cycle and its clock rate (e.g., MOPS).

Table 5.2: Estimated Operations for Common Waveform Components

Module	Parameters	Operations(Arithmetic, Logic, Multiplications, Memory)
(Real) Filtering	N = filter length	$6N + 3$
FFT	N = block length r = radix	$21N \log_r(N) + 10N - 4$
Correlation	N = block length	$6N + 3$
CRC	N = block length r = polynomial order	$5(N+r) + 1$

Convolutional Encoder	K = constraint length 1 / r = code rate N = block length	[3(1+Kr) +2K+5](N+K)+K+1
Viterbi Decoder	K = constraint length 1 / r = code rate N = block length	3+3(2K-1) + (N-4K)(4+2K+1) + [10+2K(17+3r)](N+K)
Interleaver	N = block length	5N+3
Interpolation/Decimation (CIC)	N = CIC stages R=I/D factor	3(N+Nr)+1+r
Transcendental (LUT)	N = iterations	1
Transcendental (CORDIC)	N =stages	12N + 1
Transcendental (Series)	N =block length	5N + 1
Equalizer (LMS complex)	N =block length	30N + 16

However, this metric would also be misleading because as illustrated in Table 5.3, not all types operations are needed in equal proportion, nor will different DSPs be capable of supplying the same operations at the same ratios. Further, the variance of circuitry across processors means that different algorithms may be optimal for different processors, which means a DSP's MOPS estimate can be misleading.

Table 5.3: Tabularized FFT Operations (N = number of points in FFT, r = radix)

Arithmetic (Butterfly) Operations	
Additions = (N/2) $\log_r N$ complex additions	4 $N \log_r N$
Multiplications = (N/2) $\log_r N$ complex multiplications	2 $N \log_r N$
Linear Memory Access (1 complex twiddle read, 2 complex data read, 2 complex data write)	10 $N \log_r N$
Bit Reversal Shuffling	
Linear Access (indices, read, write)	3 N
Control (Add, compare)	2 N+1
FFT Control	
Stage Loop	
Loop Control (Add, compare)	2 $\log_r N$
Other ALU operations	5 $\log_r N$
Butterfly Control Loop	
Loop Control (Add, compare)	2 (N-1)
Other ALU operations	3 (N-1)
Inner Butterfly Loop (executes $N \log_r N$ times)	
Loop Control (Add, compare)	2 (N-1) $\log_r N$
Other ALU operations	3 (N-1) $\log_r N$
Total Estimated Operations	(19 N +2) $\log_r N$ + 10N -4

Because of this interplay between waveform (or application) and processor architecture, some potentially more meaningful metrics are defined by measuring the execution time of a "basket" (or benchmark) of commonly used functions (e.g., filters, FFTs, and convolutional decoding) as optimized for each individual processor. In this approach, the

execution time for each parameterized component in the basket is measured, and converted these into some score (generally, where a higher score means less time required to execute the basket). Examples of such an approach include the following:

- BDTi score, illustrated in Figure 5.3, where the basket is defined by Berkeley Design Technology, Incorporated and focused on DSP applications,
- Dhrystones, and the later SPECint suite, intended to measure typical integer GPP operations.[6]

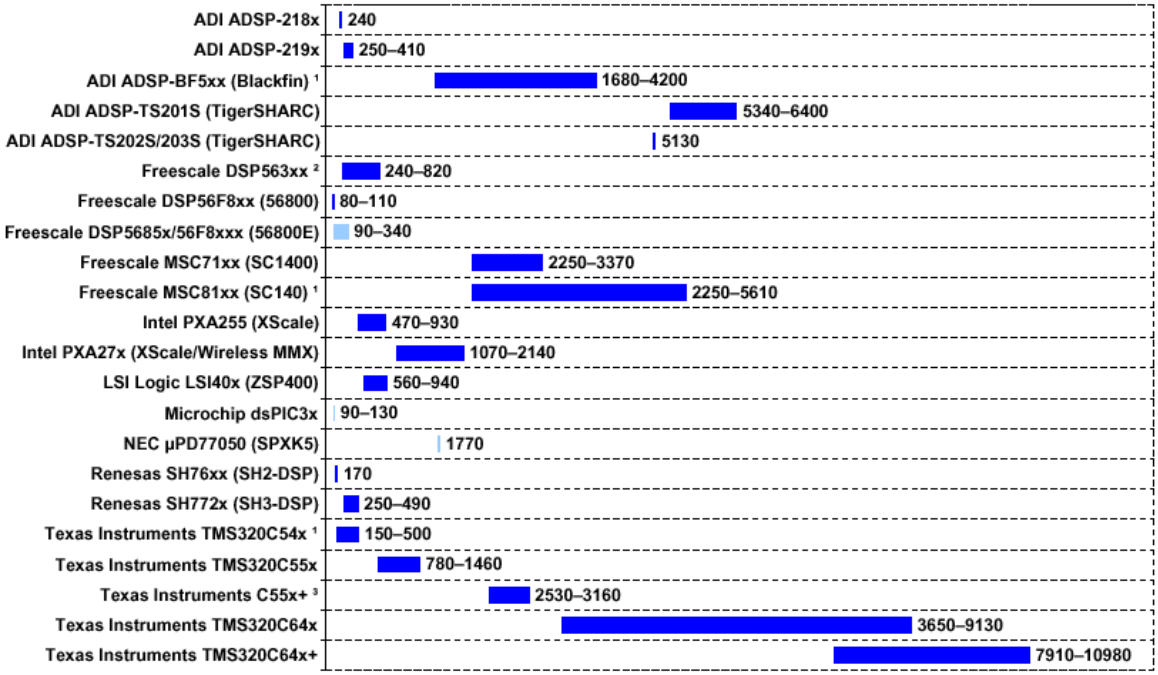


Figure 5.3: Sample BDTi score from 2006. [BDTI_06]

However, these basket-based approaches are difficult to apply to specific processors whose specific scores have not been calculated. Further, this basket-based approach can be misleading if the makeup of the testing basket or the relative frequency of occurrence of each component in the basket differs significantly from the algorithms intended for deployment. These differences are frequently exploited to lead to competing claims of superior chip designs based on the use of different benchmarks. For example, there has been an ongoing back-and-forth between FPGA vendors Xilinx and Altera about whose processors yield better performance [Maxfield_08] where the choices of functions and parameterizations for the benchmarks lead to radically different results. For instance, in [Altera_03] Altera compared the Stratix FPGA (an Altera part) with the Virtex2 FPGAs

[6] While not DSP focused, Dhrystones are frequently reported by ARM and the derivative SPECint suite is sometimes used by Intel.

(made by Xilinx) and found that the Altera parts were 10% more efficient in their use of logic elements (LE) as illustrated in Figure 5.4. In comparison, a Xilinx benchmarking study [Rivoallon_02] concluded that the same Xilinx parts were 25% more efficient than the same Altera part.

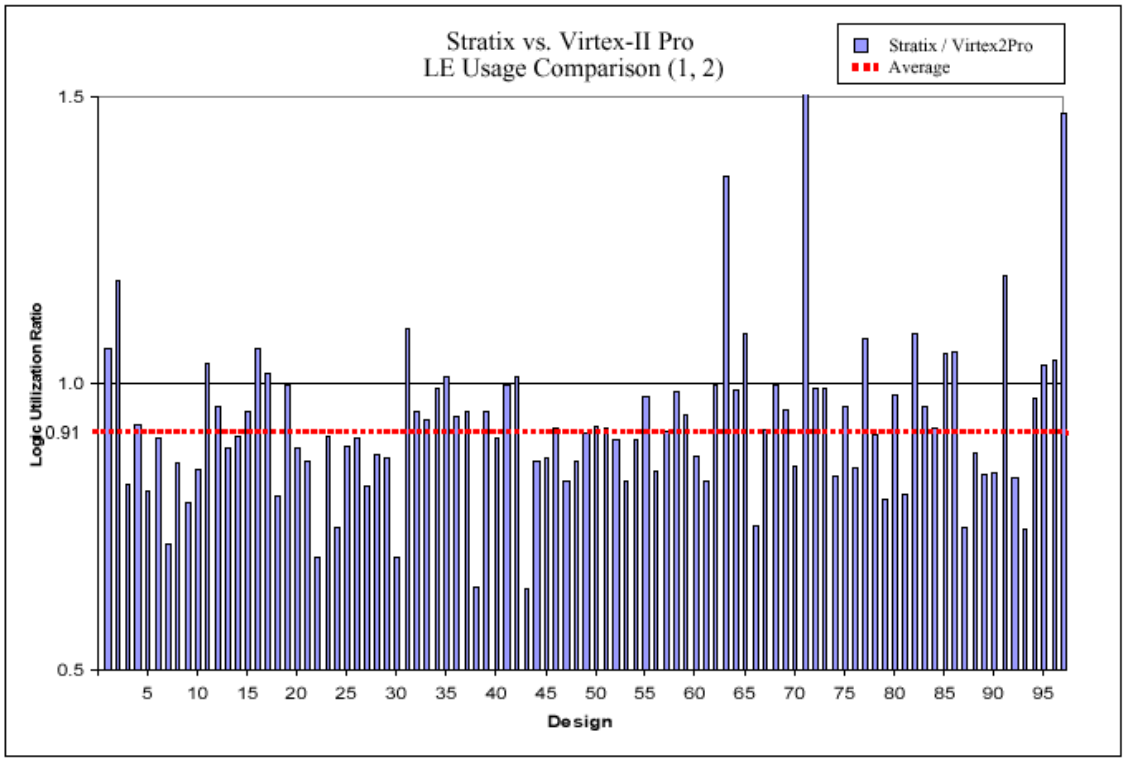


Figure 5.4: An Altera comparison of Xilinx and Altera FPGAs [Altera_03]

While they may be useful for internal consistency, it is clear that benchmarks can be misleading unless they measure exactly the set of processes that will be implemented. Accordingly, others have proposed automated methods for estimating the required execution time for parameterized waveform components [Neel_08]. While useful in the processor selection phase, such waveform specific metrics are also not well suited for a generalized study of the tradeoffs and fundamental limits for DSPs because they do not yield simple generalizable numbers for comparisons across all possible SDR implementations.

As every generalizable metric will necessarily be misleading in some way, this study will use MOPS (and MFLOPS) and MACs for our measures of processor computational capacity because of their simplicity and compactness. A **MAC** is a **M**ultiply-and-**AC**cumulate operation that can be implemented as a single instruction (e.g., TMS320C55xx) or as multiple instructions (e.g., TMS320C62xx). A MAC operation is perhaps the most commonly used operation in a DSP as it forms the basis of all filters.

Thus via the combination of MOPS and MACS we hope to give a readily-comprehendible number that is not specifically tied to a particular choice of waveforms though still somewhat significant.

5.1.1.2 Power Consumption

A microprocessor is effectively a repurposable collection of transistors. The power consumption of a transistor is typically expressed as the sum of the following two components[7]:

- *Dynamic power consumption*, P_D, which is related to the charging and discharging of the transistor (typically as applied to the output gate capacitor).
- *Static power consumption*, P_S, that occurs via a combination of tunneling and sub-threshold conduction across transistor insulators.

Dynamic power consumption can be approximated by

$$P_D \sim \alpha C v^2 f \tag{0.1}$$

where C is the effective capacitance, v is the supply voltage, f is the transistor clock rate, and α is the activity level of the transistor. Additionally, a processor will generally consume nonnegligible amounts of power driving the I/O lines to the chip. Thus there are three broad categories of power consumption for a processing chip: dynamic power consumption, static power consumption, and I/O (input/output) power consumption.

Traditionally, processor power consumption was dominated by dynamic power and I/O power considerations, but as transistor feature sizes shrank, leakage current became more noticeable such that static power consumption became a significant portion of total power consumption. For instance, Figure 5.5 shows the power consumption attributable to dynamic power (labeled as "Active") and static power (labeled as "Leakage") for various Intel processors. Note that for most of the chart, static power is negligible until this decade when static power consumption began to approach dynamic power consumption. A similar effect is seen in other processing platforms like FPGAs;[Altera_05] found that, as averaged over 99 designs, 67% of total chip power consumption was dynamic, 21.7% was static power, and 11.2% was I/O power.

[7] Transistor power consumption is also sometimes broken out to include *short circuit power consumption*, P_{SC}, which occurs as the transistor's input line changes voltage temporarily inducing a path from the supply voltage (V_{DD}) to ground. This is frequently significantly less than the other two components and ignored for high level analyses – a practice we adopt in this study.

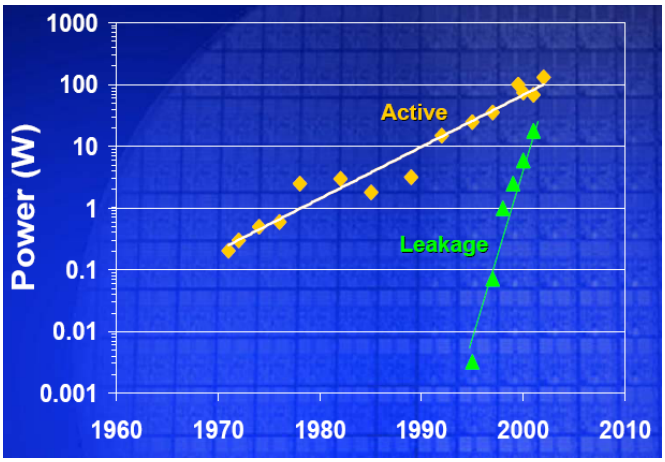


Figure 5.5: Growth in leakage currents were rapidly leading to situations where static power consumption was on par with dynamic (useful) power consumption. [Moore_03]

This steady increase in leakage current was an unintended result of the scaling process used with **CMOS** (**C**omplementary **M**etal **O**xide **S**emiconductors) integrated circuits that successively reduced the feature sizes of transistors. Examining the transistor model shown in Figure 5.6, a critical feature to the magnitude of leakage current is the gate-oxide thickness (t_{ox}) intended to insulate the gate voltage from the silicon substrate. As shown in Figure 5.7, decreasing t_{ox} leads to exponentially increasing leakage current.

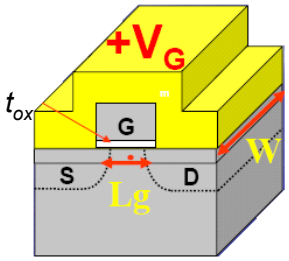


Figure 5.6: A simplified model of a transistor in an integrated circuit [Gargini_08]

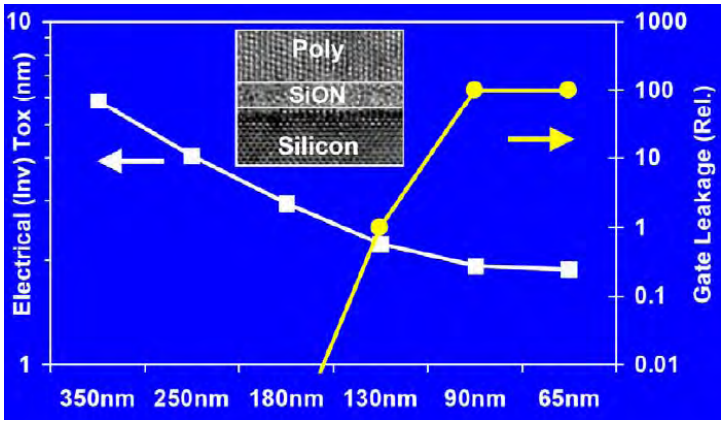


Figure 5.7: Impact of Gate Oxide Thickness on Gate Leakage Current [Gargini_08]

To combat this, researchers adopted a process where the formally Silicon Dioxide (SiO_2) gate insulator is implemented using Hafnium (frequently referred to as a High-k material in light of its high dielectric constant) to significantly increase insulation and achieve a much greater effective electrical t_{ox}. This was coupled with metal gates to decrease polysilicon depletion and achieve the desired gate capacitance and to permit further speed increases with further feature scaling as summarized in Figure 5.8.

	High-k vs. SiO_2	Benefit
Capacitance	60% greater	*Much faster transistors*
Gate dielectric leakage	> 100x reduction	*Far cooler*

Figure 5.8: Summary of benefits of material shift to High-k materials in CMOS [Gargini_08]

5.1.1.3 Processor Architectures

There is significant variation between classes of processors (e.g., GPP, DSP, and FPGA) and within classes of processors, particularly in terms of how they manage and implement their computations. A microprocessor interprets instructions stored in memory to perform calculations on inputs and data stored in memory. The results of these calculations may be stored in memory, output, or used to modify the flow of execution. The traditional microprocessor consists of at least the following components:

- A single functional unit, typically an arithmetic logic unit (ALU), that can be rapidly programmed to execute one of a number of predefined instructions
- Memory which holds instructions and data
- Circuitry for fetching, decoding, and dispatching instructions to the functional unit
- Input and output (I/O) circuitry

In the traditional microprocessor design, these components are organized according to the von Neumann architecture shown in Figure 5.9. Notice that a common bus and common memory are used for both program information and data. Modern microprocessors have evolved beyond the von Neumann architecture and many now include multiple functional units and specialized coprocessors.

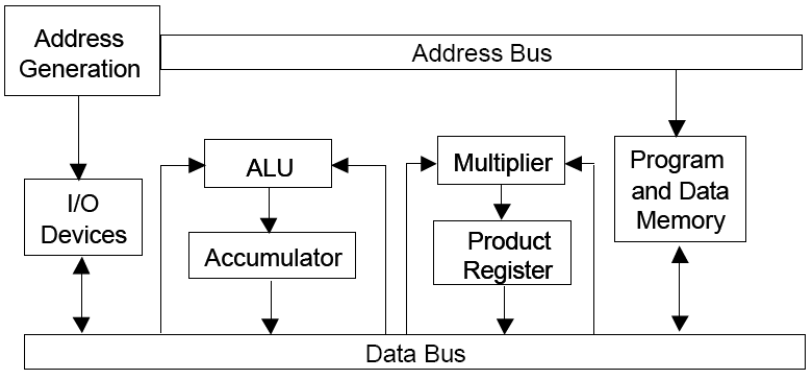


Figure 5.9: von Neumann Architecture [Neel_02]

A **D**igital **S**ignal **P**rocessor (**DSP**) is an attempt to partially optimize a microprocessor for digital signal processing applications. One way in which this has traditionally been performed is by employing the Harvard architecture shown in Figure 5.10. In the Harvard architecture, three different independent memory spaces are established and three different data and address busses used – one for program information and two for data, which are labeled *X* and *Y* in Figure 5.10. This is especially useful for the MAC operation, which is fundamental to any filter implementation, in which two variables need to be simultaneously loaded, multiplied, and the result accumulated. In the von Neumann architecture, the shared memory space and shared busses for program and data information only permit a single element to be loaded at a time – one variable or one instruction. Since two variables and one instruction (for the ALU) must be dispatched, the von Neumann architecture requires at least three sequential memory accesses to perform one leg of a MAC operation.

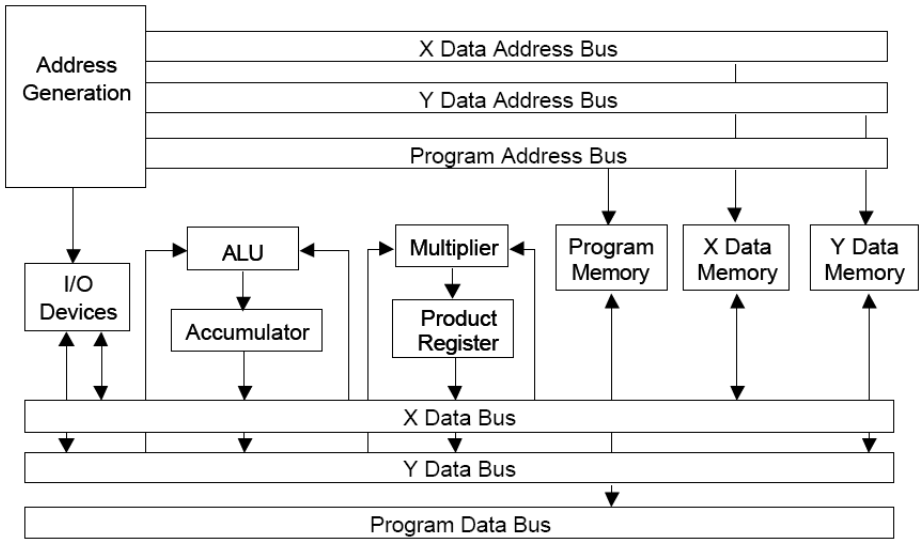


Figure 5.10: Harvard Architecture [Neel_02]

In practice, the generalized Harvard Architecture has been modified to include many new features designed to enhance signal processing performance. Some of these features include deeper pipelines, specialized addressing modes, such as bit-reversed addressing to facilitate the FFT, **V**ery-**L**ong **I**nstruction **W**ord (**VLIW**), and **S**ingle **I**nstruction **M**ultiple **D**ata (**SIMD**). In architectures that utilize VLIW, the DSP is capable of fetching multiple instructions at a time that are executed simultaneously in parallel. Architectures that use SIMD have functional units that are capable of treating its operands as a number of smaller independent operands so that a functional unit performs the operation indicated by the instruction on the smaller independent operands. For instance in a DSP that can exploit SIMD, two 32 bit integers can be operated on as four 16-bit words in the same period of time.

Additionally, further optimizations include specialized arithmetic rules (e.g., Galois Field arithmetic), numerous co-processors (e.g., a Viterbi co-processor), advanced looping support (e.g, zero-overhead loop or block repeats), and inclusion of circuitry that permits multiple operations to be executed in a single cycle by a single instruction (common examples of which are listed in Table 5.4). Additionally, many processors will include dedicated circuitry for interfacing with expected I/O formats and for implementing peripheral operations that are expected to be useful. Further optimizations are possible in terms of the size and management of memory (e.g., L1 vs L2 vs off-chip memory) and the processes by which memory is moved around within a chip. Many of these features are exhibited by the DSP architecture shown in Figure 5.11 wherein dual data memory access is enabled by the data (.D) units with separate access to program memory (ala a Harvard architecture) and the eight functional units permit VLIW operation with most of these units supporting SIMD and other specialized instructions. Additional computational capacity is provided by Viterbi and Turbo co-processors (labeled VCP and TCP) especially designed for 3G convolutional codes.

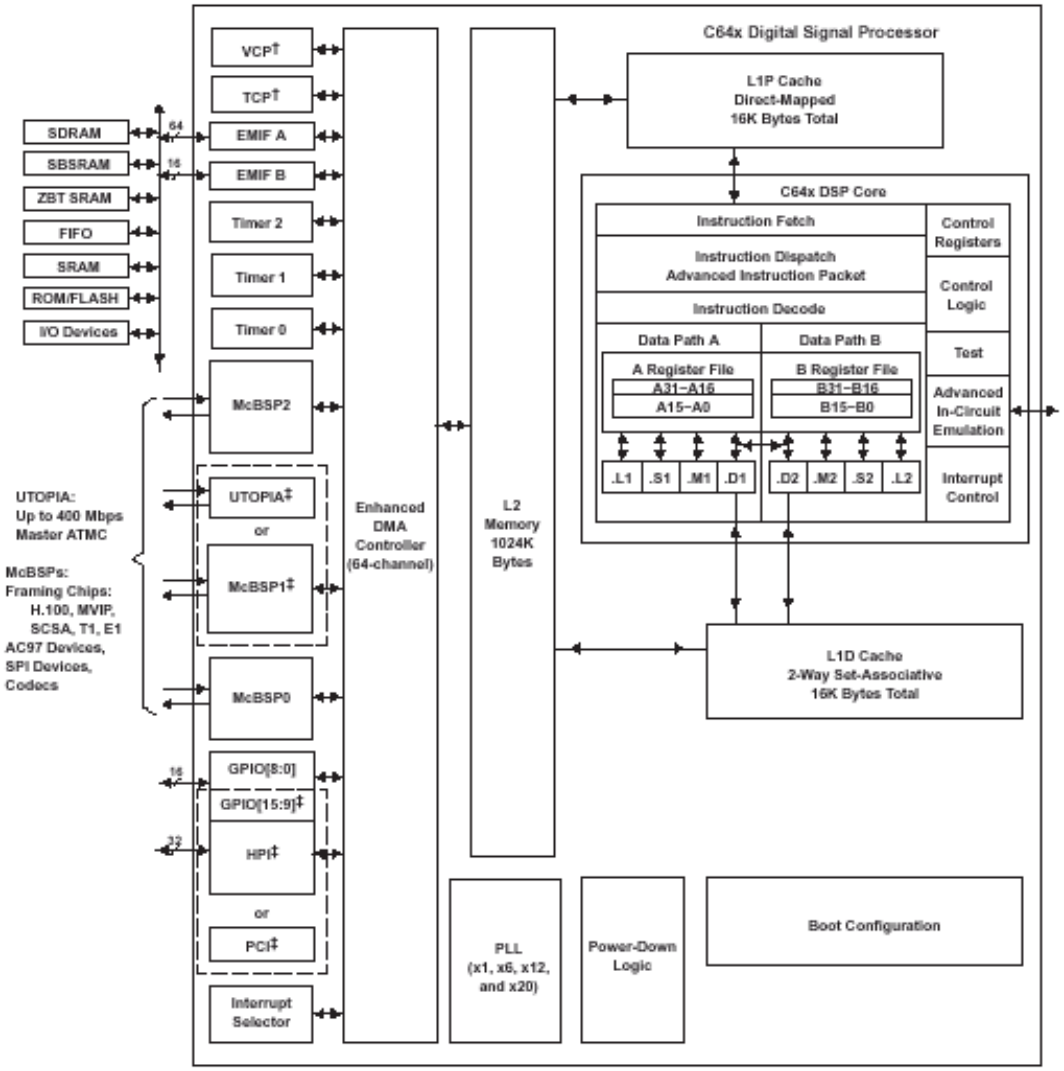


Figure 5.11: Texas Instrument's TMS3206416T Architecture [TI_08]

Table 5.4: Examples of Common Multi-operation single-cycle instructions

Specialized Instruction	Operations
ABSALU	A typical ALU instruction (ADD, SUB) is combined with an absolute value operation. Useful for distance calculations and L1 norms.
ADDSUB	The ALU can simultaneously add and subtract two numbers and/or add/subtract two numbers of the native precision
AVG	The processor implements (A+B)/2
BDEC	The processor decrements a counter and branches
BPOS	The processor branches on general conditions, e.g., R>0
BITR	The processor reverses the bits of a register in a single cycle

GMPY	The processor supports Galois Field arithmetic (useful in some error correction codes).
MAC	A single-instruction (functional unit) multiplication and accumulation. Note that when both a multiplier and an ALU are required to implement this, a processor is not considered to exhibit the MAC modifier unless it does not support VLIW
MAX2	The process performs the MAX operation for two pairs of words of native precision.
MEM2	The data bus width of a processor is such that a single instruction fetches 2 words.
NOREG	The processor memory maps all registers so there's no need for an instruction to load registers from memory. Processors which do this, however, tend to clock much slower
SIDE_SUM	Adds up bytes in a word.
VECT	A single cycle complex multiplication. Note that several DSPs have an instruction which implements a complex multiplication (or LMS update or an FIR cycle) but these are multi-cycle instructions. Here, we're only modeling situations where x_1 * y_1 + x_2*y_2, x_3*y_3 + x_4*y_4.
VSL	A process by which a register is shifted left and an input 1 or 0 is appended to the right most bit. Useful for keeping track of paths (saves an instruction).

Note that many GPPs incorporate these same features (though generally not communications-specific coprocessors such as a Turbo decoder coprocessor) and will frequently include other features that are useful in more general settings. An example of this is branch prediction circuitry wherein the chip remembers past executions of an instruction and uses this memory to load the most likely set of branched-to instructions into the pipeline before the branch condition finished evaluating.

Also note that while the architecture in Figure 5.11 only depicts a single computational core, as discussed in the Section 5.2.1.5, both GPPs and DSPs have begun to incorporate multiple cores onto a single chip. In GPPs this is typically done via replications of the basic GPP core (e.g., the AMD Athlon dual-core) while DSPs tend to (but not exclusively) incorporate ARM cores to support other tasks frequently found on platforms (e.g., video support or user input control). This allows a processing platform to leverage the additional computational capacity of additional computational cores without increasing the clock speed and without incurring the added I/O power consumption while speeding up communications between the cores, typically via shared memory spaces.

Field **P**rogrammable **G**ate **A**rrays (**FPGA**s) adopt a radically different computational architecture wherein instruction handling is eschewed in favor of a sea (or matrix) of programmable logic elements connected via a number of local and longer distance (frequently "global" as in the entire chip) connections. Generally each of these logic elements operate on 1-bit to 4-bit words that can be combined together to provide effectively arbitrary word-widths. Programming these elements is performed by changing the bit values stored in the memory elements that determine the operation of the logic element's components.

In Xilinx parts, this logic element is called a *Slice* and includes two 4-input function generators, carry logic, arithmetic logic gates, wide function multiplexers and two storage elements. Each four-input function generator is programmable as a 4-input LUT, as

distributed random-access memory, or as 16-bit variable-tap shift register element. These slices are grouped together in units called *Configurable Logic Blocks* (CLB) and can be connected to the various local and global connections via a Switch Matrix as shown in Figure 5.13. These CLBs are then arrayed in a broader matrix of connections as shown in Figure 5.14.

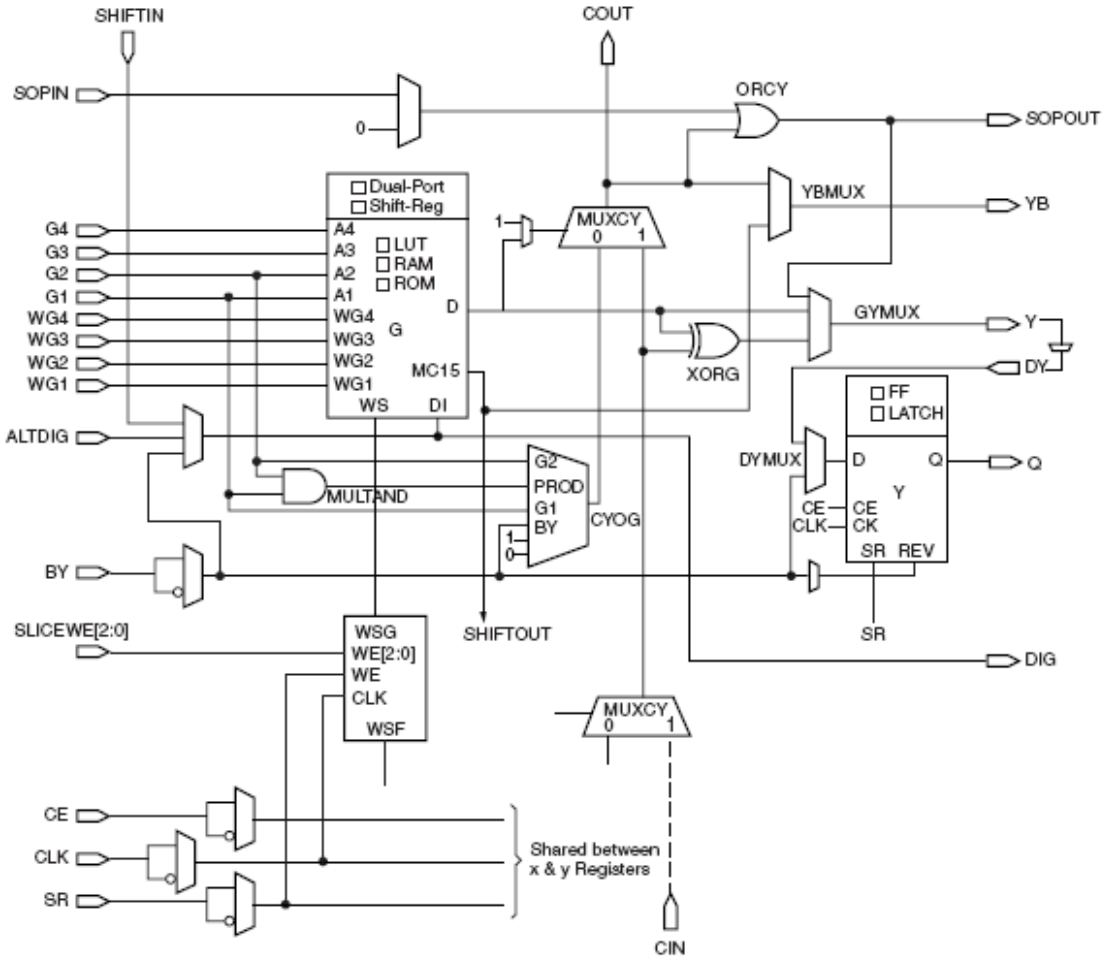


Figure 5.12: A Virtex II Slice [Xilinx_07]

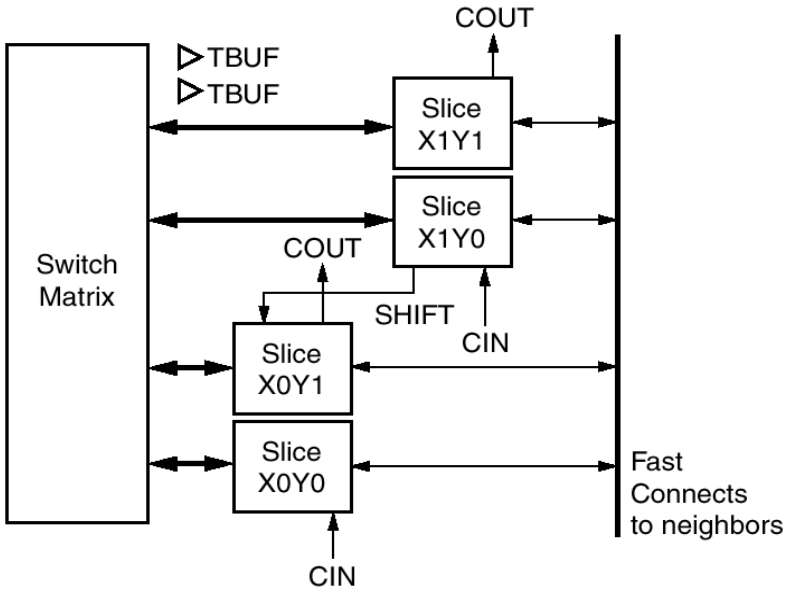


Figure 5.13: A Xilinx Virtex II CLB. [Xilinx_07]

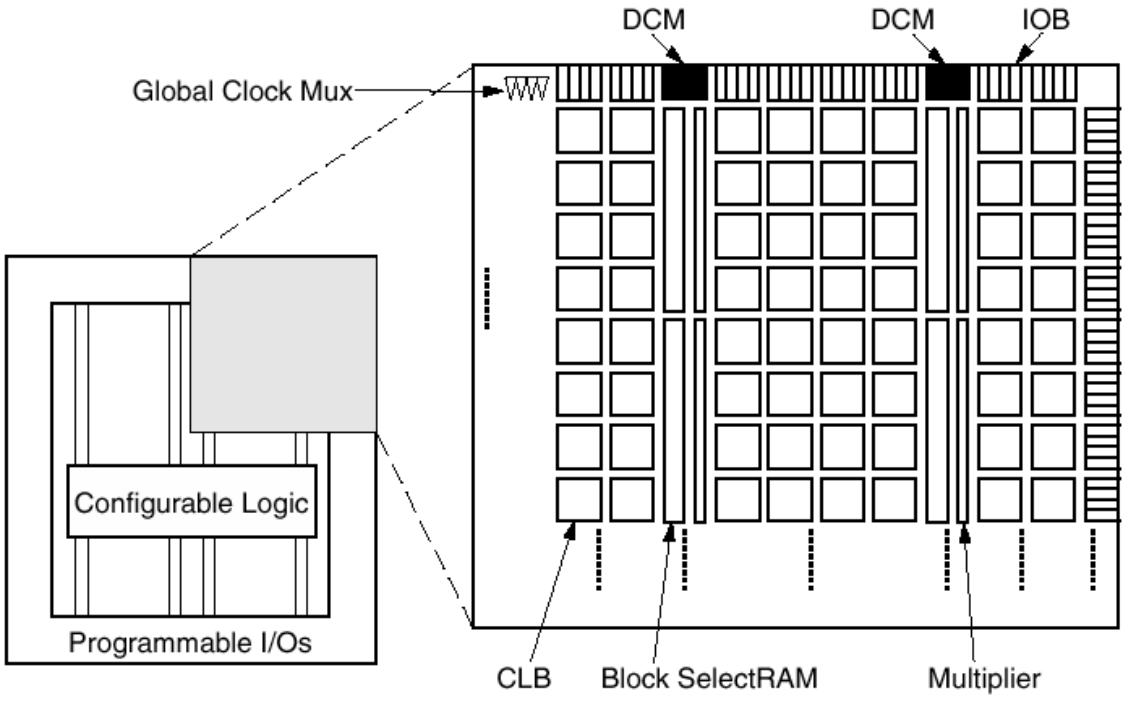


Figure 5.14: Virtex II Architecture Overview [Xilinx_07]

In general such an architecture provides a tremendous amount of computational resources, though it does have the following drawbacks:

- Requires a long time to program (and thus reconfigure) – normally on the order of milliseconds [Neel_02]
- Implements operations that span slices much less efficiently than embedded silicon (e.g., multipliers)
- Is not well-suited for applications with significant amounts of branches (e.g, control applications).

To address the first issue, many FPGAs have begun to support partial reconfiguration of the chip so that some portions can be modified [Silva_06] while other portions continue to run. The latter two issues are solved in the same manner – by embedding the relevant silicon in the logic element matrix. For example this includes the embedded 18x18 multipliers shown in Figure 5.14 (realized as "DSP blocks" in Stratix II components for embedded complex multiplications), embedded memory, and embedded co-processors, particularly embedded microprocessors. In general this arrangement allows FPGAs to achieve much greater computational capacities, but typically with greater power consumption (due to high static power consumption levels) than DSPs, though typically much less power consumption than a GPP. As a tradeoff, however, FPGAs are generally considered the most difficult platform to program and have other practical issues related to the dynamic management of bit images (the means by which FPGAs are programmed).

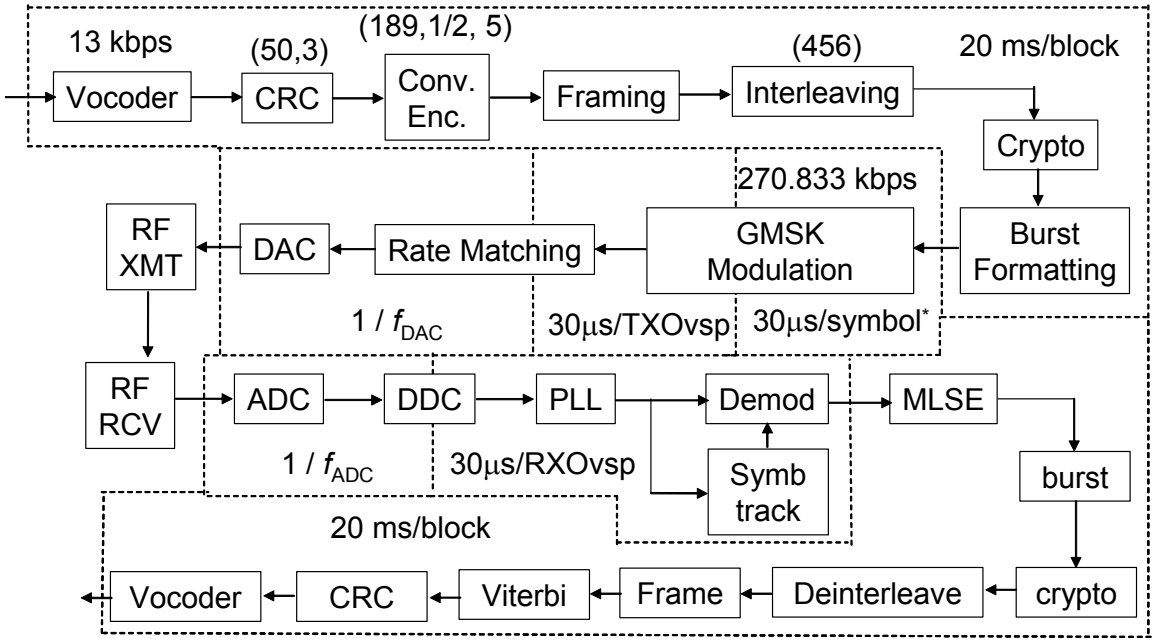
5.1.2 Relationship with Traditional SDR Performance Requirements

Fundamentally, the computational capacity and power consumption of DSPs contribute to the power consumption of the SDR as well as determining which waveforms (or applications or processes) the SDR can support. To a great extent, computational capacity determines the degree to which an SDR is "future proof." To illustrate these relationships and the influence of processor architecture, the following provides an extended analysis of implementation of GSM and 802.11a waveforms on DSPs based on the methods presented in [Neel_05].

5.1.2.1 A GSM Case Study

The GSM transceiver at the physical layer can be visualized as shown in Figure 5.15 where key processes have been separated into "components". In this componentization there are six distinct clock domains (sets of components subject to the same timing requirements) determined by a specific component in each domain. The vocoder processes blocks of data at a rate of 20ms/block. Supporting GSM modulation requires a minimal symbol processing rate of 30 μs/symbol. The modulation and demodulation/synchronization processes are frequently oversampled, specifying two

additional clock domains. The sampling rates of the ADC and DAC specify two further clock domains (assuming that the ADC and DAC do not have the same sampling rate and are not equal to the modulation rates). Within these clock domains, further componentization is primarily determined by logical block processing boundaries, the exception being the componentization indicated for the demodulation operation which has arbitrarily been divided into components for phase recovery, symbol timing recovery, and symbol demodulation. Each component must then operate at least as fast as the rate indicated for that clock domain.



(*) Since 8 timeslots, up to 30μs/symbol is available instead of raw 3.7 μs/symbol.
(**) MLSE constraint length 5 is a typical value.

Figure 5.15: Key Processes in a GSM Transceiver at the Physical Layer [Neel_05]

To parameterize this model we are formally assuming a full-rate traffic channel, a transmit oversampling factor of 4, a receive oversampling factor of 8, assuming the DAC runs at 8.7 MHz, the ADC at 17.3 MHz, and ignoring control processing (e.g., power control and handoffs) and GSM encryption. Table 5.5 summarizes the estimated number of MOPS required for each module[8] and Table 5.6 further breaks these numbers down for key subprocesses. Table 5.7 then gives the estimated number of cycles and typical power consumption numbers for implementing these processes on several different DSPs[9] and

[8] This is relatively close to Crocket's ball-park estimate for GSM of 100 MIPS in [Crocket_98].
[9] Typical numbers allow us to ignore more complicated processes such as dynamically invoking sleep modes or controlling DSP voltage and/or clocks to reduce power consumption. In practice, power consumption should be lower as all of the surveyed DSPs were easily able to implement the identified processes.

Table 5.8 gives estimated metrics for implementation on selected FPGAs.[10] **In this case, both the DSPs and the FPGAs had sufficient computational capacity to implement the waveform (GSM), but the FPGAs consumed noticeably more power**.

Table 5.5: Estimated MOPS for key GSM components

Module	Estimated MOPS
Digital Downconversion	28.4
Coherent Carrier Recovery	21.9
Digital Upconversion	12.6
Equalizer (5 MLSE)	10.0
Channel Coding	8.4
Other	5.2
GMSK Modulation	2.9
GMSK Decision	0.5
Total	**89.9**

Table 5.6: Estimated MOPS for key subprocesses

Module	MOPS
FM Phase Shift Filter	14.4
Complex Multiply (DDC)	13.0
Interpolation (Upconversion)	12.6
MLSE (5 tap equalizer)	10.0
Decimation (DDC)	8.9
Viterbi (FEC)	7.8
Sine wave gen. (DDC)	6.5
GSM Vocoder	5.0
Atan2 (FM PLL)	3.2
Gaussian Filter (Polyphase)	2.5
Total	83.9

[10] The Stratix II uses DSP blocks which have four embedded multipliers and **A**daptive **L**ogic **M**odules (**ALM**) which is approximately equivalent to a Virtex slice.

Table 5.7: Estimated Metrics for Implementing on Selected DSPs

DSP	Cycle (MHz)	Power (mW)	Excess Cycles (MHz)
Blackfin	400	466	356.4
Blackfin	750	875	706.4
CEVA-x1620	450	300	405.9
SC140	300	200	233.8
TI VC5441	133	90	79.2
TI VC5501	300	110	230.9
TI C64xx	300	250	254.9
TI C6416T	1,000	1,650	954.9
TI C67xx	100	500	50.4
TI C67xx	255	1,400	205.4
TS-203	500	2,170	444.4
ZSP-540	350	105	297.7

Table 5.8: Estimations of Metrics for Implementing GSM on Selected FPGAs

FPGA	Slices/ALM	Multipliers/ DSP Blocks	Block RAMs	Static Power	Dynamic Power	Total Power (mW)
Virtex II	6460	18	70	353 mW	2295 mW	2648
Virtex IV	6460	18	70	300 mW	780 mW	1080
Stratix II	5943	5	70	418 mW	450 mW	868

5.1.2.2 An 802.11a Case Study

The 802.11a transceiver at the physical layer can be visualized as shown in Figure 5.16 where key processes have been separated into "components". In this case, the timing requirement is for recovery of a symbol within 4 μs to allow proper operation of the CSMA/CA function, which makes the receiver chain the limiting factor for computational capacity. When recovering the preamble, both coarse and fine synchronization will need to be performed based on pilot symbols, and then when recovering data symbols the following steps will need to be performed: an FFT, direct digital down conversion, channel estimation and equalization, descrambling, a Viterbi decoder, and a cyclic redundancy check.

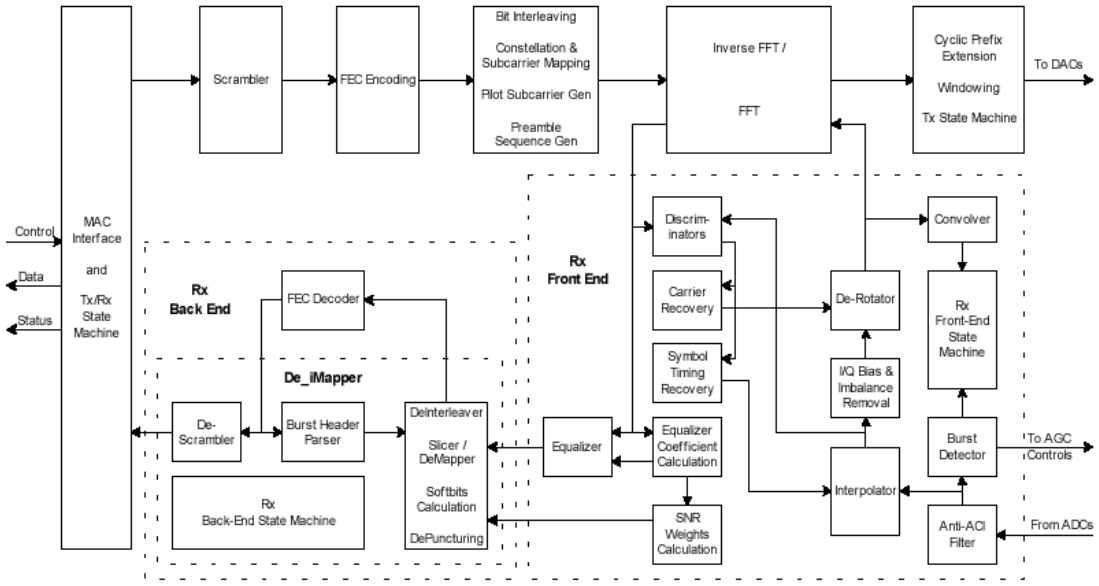


Figure 5.16: Key processes in 802.11a physical layer [Commstack_03]

For these processes, Table 5.9 gives the estimated number of cycles required to implement these processes on several different DSPs. Note that the total column excludes the preamble (shaded in blue), items in red violate the 4 μs symbol timing requirement, and the parenthetical estimate on the TMS320C6416T Viterbi estimate if the Viterbi coprocessor (VCP) is used. Note that none of these DSPs had the computational capacity to execute all of the required functions in the required time. However, it is possible to identify FPGAs that can support the requisite functions as shown in Table 5.10, it would be possible to implement these processes on FPGAs though with significant dynamic power consumption.

Table 5.9: Estimated times required to implement key 802.11a receiver physical layer processes

	Coarse	Fine	FFT	DDC	Channel Est.	Channel Eq.	Descr.	Viterbi	CRC	Total
Blackfin	2.6	6.6	3.2	26	0.04	0.36	0.29	75	0.3	105.19
CEVA	2.1	4.3	3.8	44	0.07	0.3	0.48	62	0.5	111.15
SC140	9.3	16.5	9.1	67	0.18	1.5	0.62	328	1.9	408.3
MSC8216	**1.53**	**2.7**	**1.5**	**20**	**0.03**	**0.25**	**0.1**	**54**	**0.31**	76.19
C54	22	36	16	47	0.1	3.35	1.35	360	2.8	430.6
C55	20	28	21	100	1.6	2.7	1.1	950	5	1081.4
C6416T	**1.15**	**1.92**	**1.34**	**18**	**0.02**	**0.13**	**0.2**	**15 (11)**	**0.45**	20.14
C67	7	10	10	67	0.1	1.2	0.7	285	1.5	365.5
TS-203	**3.8**	**2.3**	**1.78**	**20**	**0.13**	**0.27**	**0.4**	**13.7**	**1.3**	37.58
ZSP-540	9.3	16	9.1	57	0.18	1.5	0.6	328	1.9	398.28

Table 5.10: Estimated resources and dynamic power consumption estimates for implementation of 802.11a components on selected FPGAs

	Virtex Slices	Virtex Multipliers	RAM Blocks	V2 Power (mW)	V4 Power (mW)	Stratix II Power (mW)
FFT	1331	9	8	1051	417	245
Coarse	462	12	0	494	169	85
Fine	1100	2	0	739	313	202
Channel Est	776	10	2	673	250	145
Channel Eq	40	4	1	93	25	8
Viterbi	1084	0	2	701	304	202
DDC	805	0	2	528	228	150
			TOTAL	4279	1706	1037

5.2 Trends in DSP Performance

To analyze trends in DSP performance, an Excel database was constructed from a survey of the parts available from the following manufacturers: Texas Instruments (TI), Analog Devices (ADSP), and Freescale. To construct the DSP database, every DSP datasheet posted on each surveyed manufacturer's website was downloaded and collected the data in the following categories for each DSP: part #, clock speed, number of cores, peak # operations per cycle, peak number of MAC operations per cycle, typical core voltage, typical core current, core power consumption, bit field-width, numeric representation, year of first manufacture, and fabrication process. To augment this trend analysis, reports from Intel and the International Technology Roadmap for Semiconductors (ITRS) were reviewed to facilitate discussion of trends in GPPs and transistors.

5.2.1 Forces Driving DSP Trends

Commercially parts are largely driven by the needs of the market into which they are sold and improvements in underlying technologies, in this case the transistor fabrication technologies. For processors, the largest commercial market are the GPP market which has somewhat recently become concerned with power consumption (and more specifically power density) and has moved towards multi-core solutions to allow computational capacity to continue to increase without significant increases in power consumption. Likewise, DSPs have also been driven by the dominant markets demands at the time where there was a need to provide integrated control and DSP operations for handsets (and later video) at low power. For TI, for example, these needs spawned the OMAP processor lines (and later, the DaVinci).

To keep up with more general increases in waveform complexity, DSPs have continued to increase their available computational capacity by increasing the number of simultaneous operations per cycle while increasing clock rates at a slower rate. Examples of past driving waveforms include those listed in Table 5.1 with more recent designs based on the expected

deployments of LTE [TI_07] whose demands will begin to drive DSPs more strongly towards multicore architectures.

5.2.1.1 Clock Speeds

Because of the size of the GPP market and its role in our daily lives, it is fairly commonly known that to avoid significant heat dissipation issues, GPP clock speeds began to level off at the beginning of the decade and have since been fairly flat at just under 4 GHz as shown in Figure 5.17.

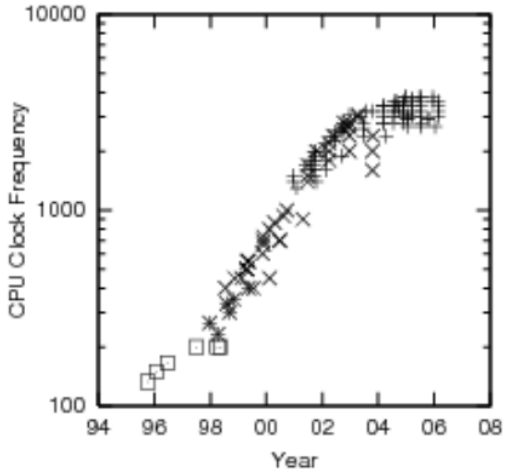


Figure 5.17: Trend in GPP Clock Speeds [McKenney_07]

Using our database of commercially available DSPs, we generated the scatter-plot of sampling rates versus year of initial manufacture shown in Figure 4.14. Unlike with GPPs, there has not been as noticeable flattening in clock speeds, though peak DSP speeds are significantly less than that of GPPs. Also, while both fixed-point and floating-point DSPs speeds have been increasing, the rate has been faster for fixed-point processors.

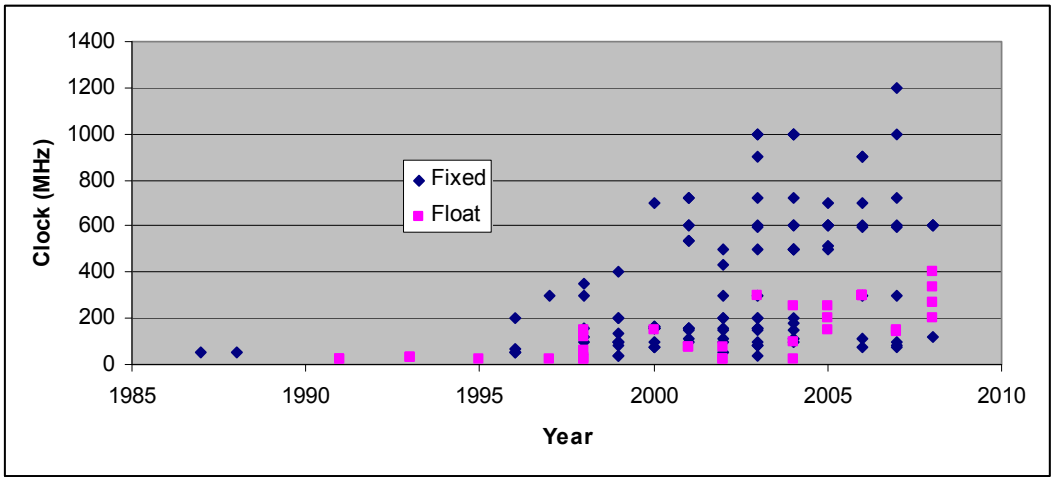


Figure 5.18: Trend in Clock Rates for Commercially Available DSPs.

5.2.1.2 Computational Capacity

Using the DSP database, we plotted the trends in **M**ega **O**perations **P**er **S**econd (MOPS) (**M**ega **FL**oating point **O**perations **P**er **S**econd – MFLOPS for floating point DSPs) and 16-bit **M**ega **M**ultiply-and **AC**cummulate per **S**econd (MMACS) as shown in Figure 5.19 and Figure 5.20. As with clock rates, the computational capacity of fixed-point processors has grown at a much faster rate than for floating-point processors. However, this relative increase in computational capacity has been greater than the relative increase in clock rates. This is evidenced by fixed point processors clocking 3 times faster than peak floating point processors (1200 MHz vs 400 MHz) but performing 6.67 times more operations than for floating point processors (16,000 MOPS versus 2,400 MFLOPS). Note that plotting the picoChip PC102 would obscure this trend as it achieves 38,400 MMACs. The PC102 is discussed in more detail in later sections.

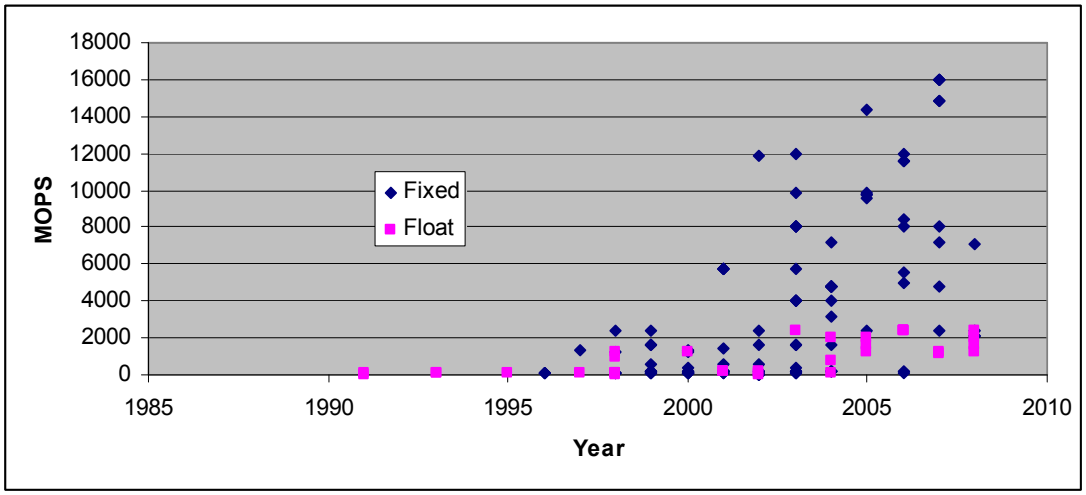


Figure 5.19: MOPS have increased significantly faster than MFLOPS

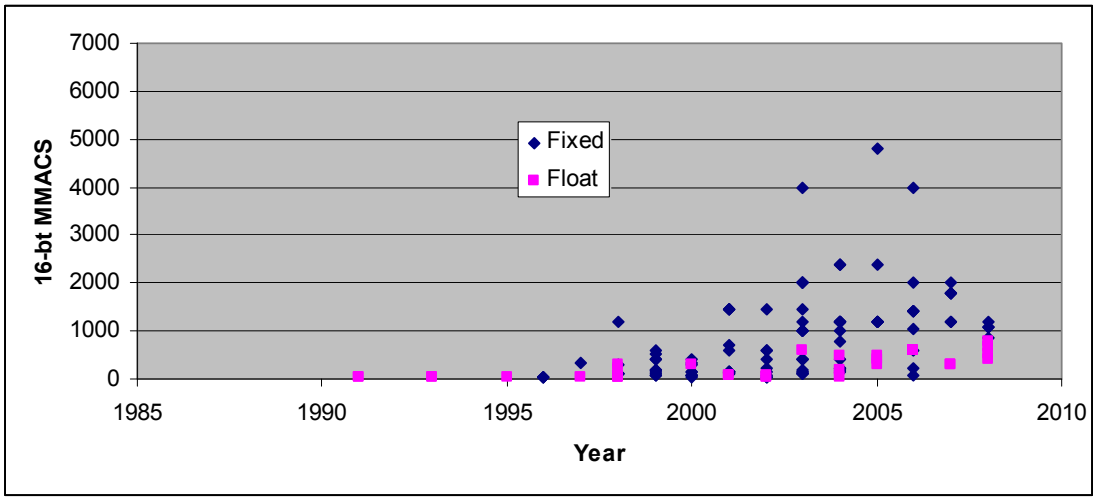


Figure 5.20: Trend in DSP MACS

5.2.1.3 Power Consumption

For GPPs, power consumption levels (as measured by thermal design power) have closely tracked increases in clock rates as shown in Figure 5.21. For DSPs, a similar relationship also holds. To somewhat reduce the variations between processor power consumption due to varying numbers and types of peripherals, the power estimates are formed as the product of typical core voltage supply levels and typical core current levels. The result of this survey is shown in Figure 5.22 and a related scatter plot between clock rate and power are shown in Figure 5.23. Note that power has also been tending upwards for DSPs, particularly for fixed-point processors, though much less so than was the case for GPPs. Also note that while power has not increased as dramatically, it is still correlated with clock speed as would be expected from equation (0.30).

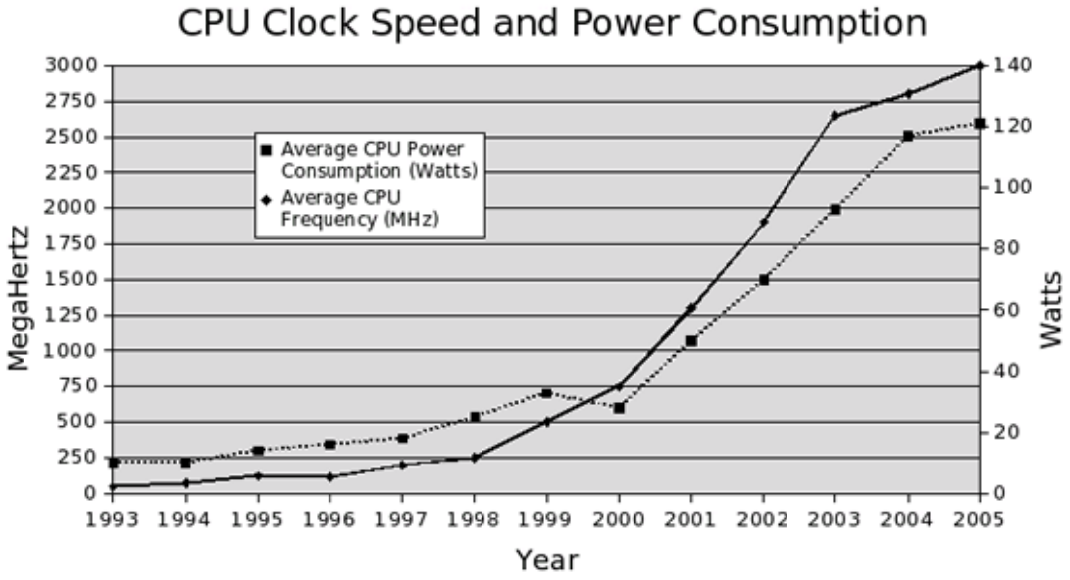


Figure 5.21: The trend lines in GPP clock speed and power consumption are highly correlated. [Linux_07]

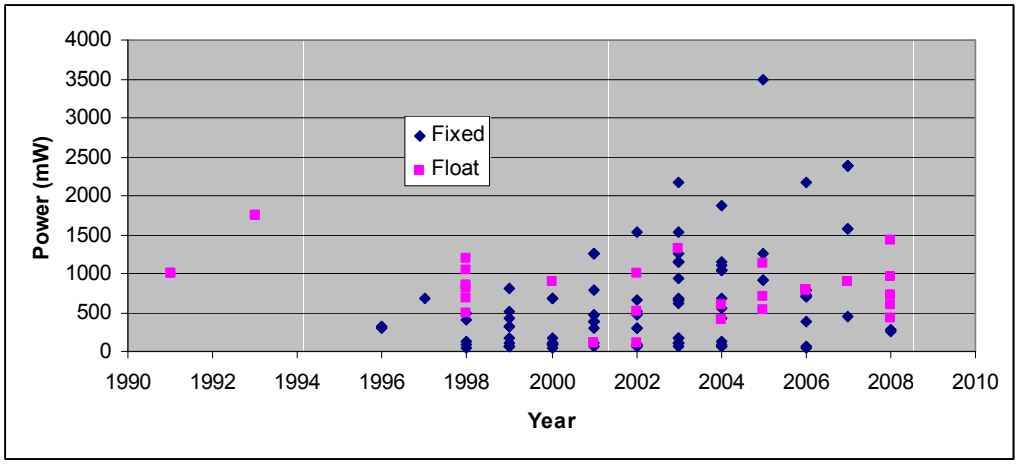


Figure 5.22: DSP Power Trends

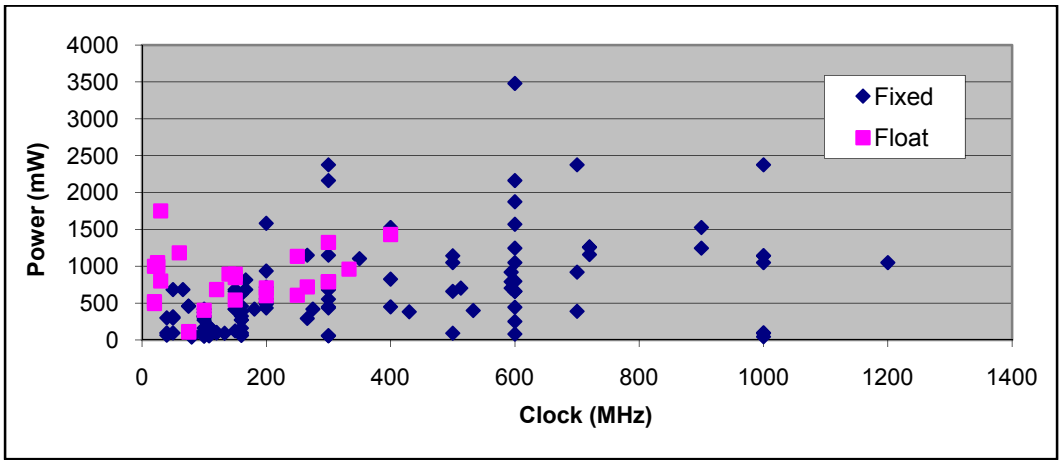


Figure 5.23: Relationship between Power and Clock Rate

5.2.1.4 Transistor Trends

While GPP clock speeds have saturated as shown in Figure 5.17, this is not due to an end to Moore's Law as the number of transistors per device have continued to increase at the same doubling rate of every 24 months as shown in Figure 5.24. Largely, this is due to continued decreases in transistor feature size as discussed in Section 5.3.1.1. While these gains do not show up in clock rates they do show up in increases in the number of cores per device and number of operations per cycle as discussed in Section 5.2.1.5.

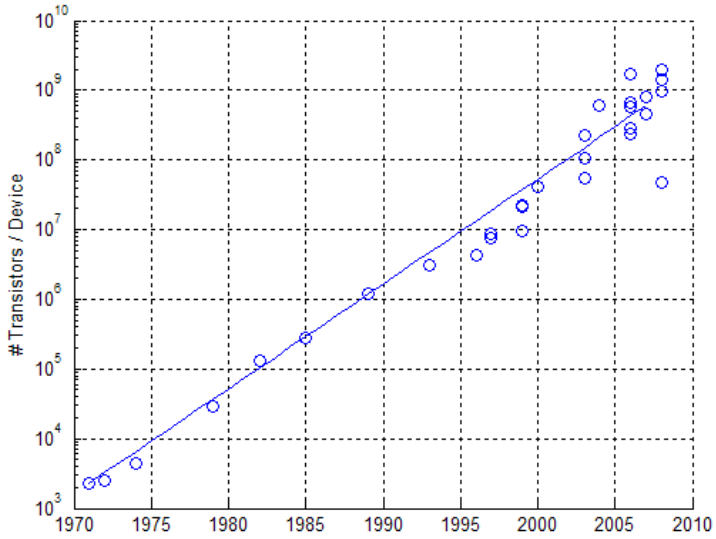


Figure 5.24: Based on data listed in [Wiki_08], the number of transistors / device has continued to double every 24 months.

5.2.1.5 Simultaneous Operations Trends

Unlike for GPPs for which there has been a clear and recent shift to multi-core architectures as evidenced by the sharp rise in spending on multi-core servers shown in Figure 5.25, multi-core DSPs have a longer history and less noticeable transition to multi-core DSPs as shown in Figure 5.26. Note that in addition to the well-known ARM + DSP combinations (e.g., OMAP and DaVinci), lower-end multicore DSPs have been available since the 90's (e.g., TMS320VC5441) though are recently becoming more popular in higher end processors (e.g., MSC8144). Beyond simply increasing the number of cores, other architectures, such as Single-Instruction-Multiple-Data (SIMD) and superscalar (e.g., VLIW), have been used to increase the number of operations that are executed per cycle as shown in Figure 5.27. Note that as fixed point operations are more amenable to SIMD operations, fixed point processors have exhibited a much larger increase in the number of operations completed per cycle. It is this trend in combination with the relative increase in clock rates that explains the divergence in MOPS and MFLOPS that was shown in Figure 5.19.

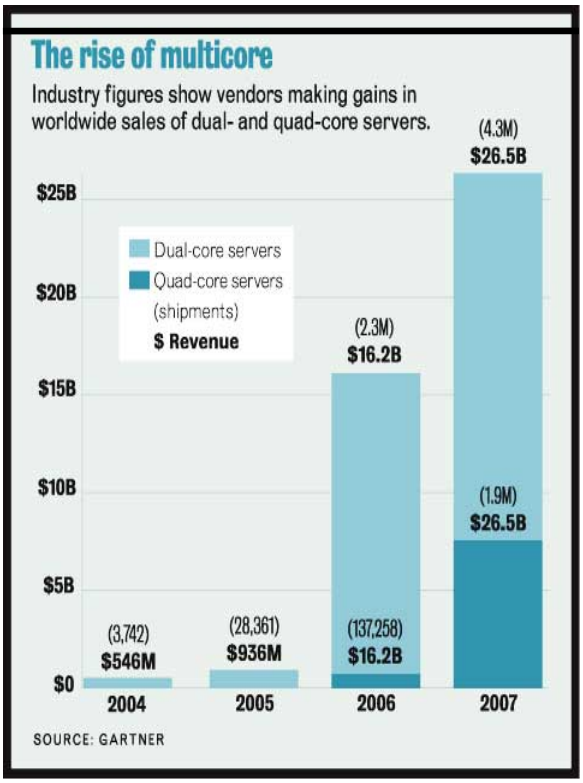


Figure 5.25: There has been a rapid uptake in multi-core processors for servers. [Brodkin_08]

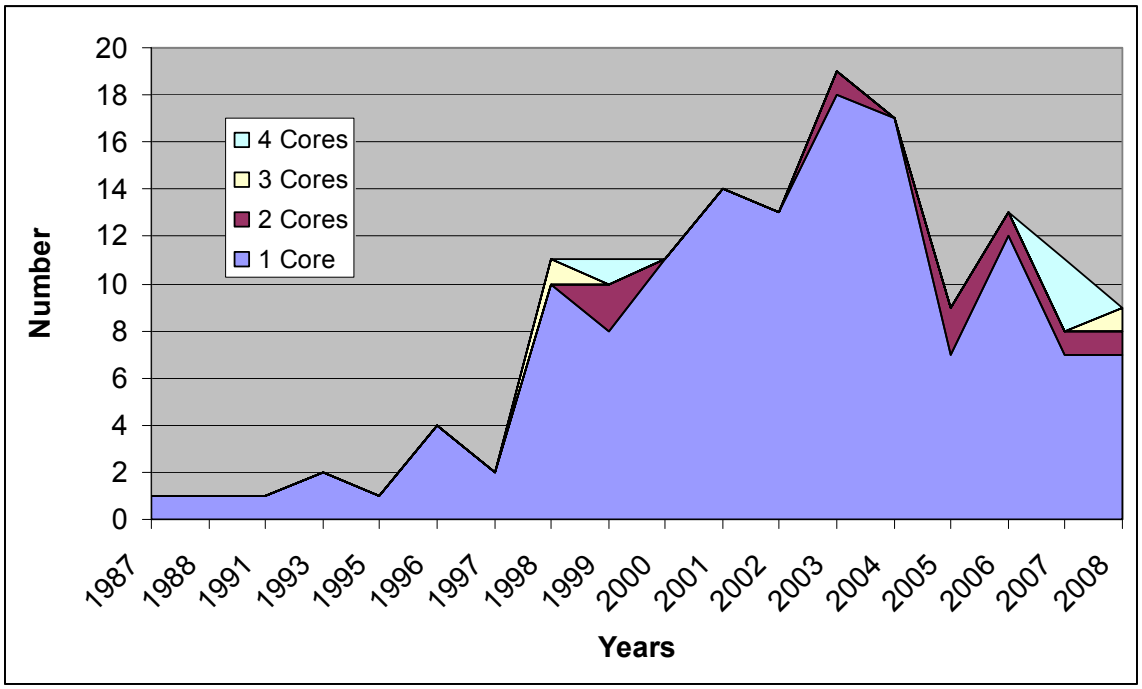


Figure 5.26: Trends in Number of Cores in Surveyed DSPs

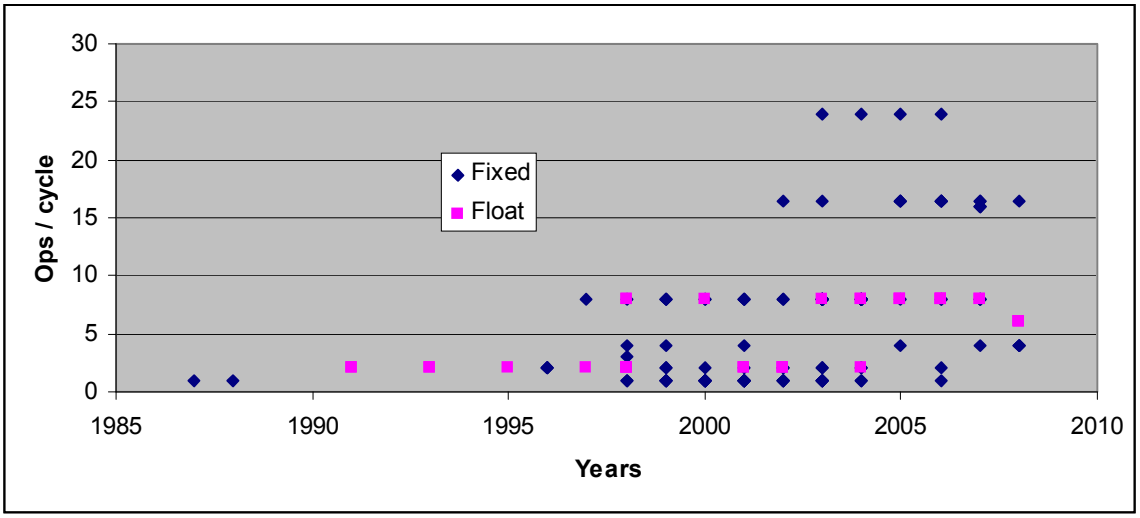


Figure 5.27: Trend in Operations / Cycle

5.2.1.6 Transistor Cost Trends

As feature sizes have continued to decrease, fabrication costs have continued to rise exponentially while revenues have flattened as shown in Figure 5.28. This makes the margins tighter on each subsequent generation of processors. At some point, these trends may so significantly limit the available capital for further retooling and research that subsequent scaling significantly slows down, a trend perhaps already in evidence as the ITRS in the delayed transition to 450 mm wafers.[11]

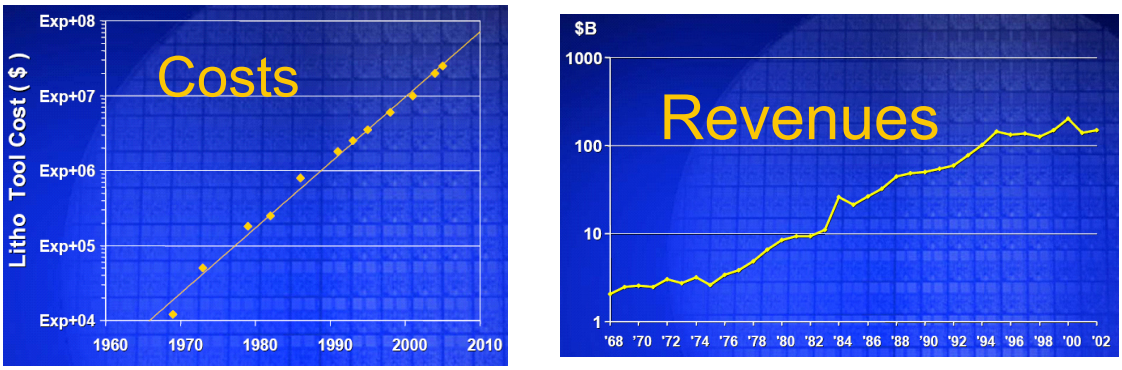


Figure 5.28: The costs for lithography tools have continued to rise exponentially while revenues have flattened. [Moore_03]

5.2.1.7 Composite Performance Trends

A useful composite metric in evaluating processors is the "bang-for-the-buck", i.e., how much power does it cost to achieve a particular computational capacity. As mentioned in Section 5, computational capacity is strongly correlated with power consumption. For years for GPPs computational capacity was a direct function of the clock rate which in turn means that power should be strongly correlated with GPP clock rate – a phenomenon exemplified in Figure 5.21. A more precise between power and performance can be derived for specific architectures as was given in [Grochowski_06] and illustrated in Figure 5.29. Note that as long as similar architectures were used (486 to Pentium 4), power was related to performance as Power = performance ^ 1.75 where performance was measured via an Intel metric and referenced to the performance of the 486. However, the shift to dual cores allowed for a dramatic movement off of this curve.

[11] This is not directly related to transistor size, but is indicative of the rising costs associated with transistor fabrication.

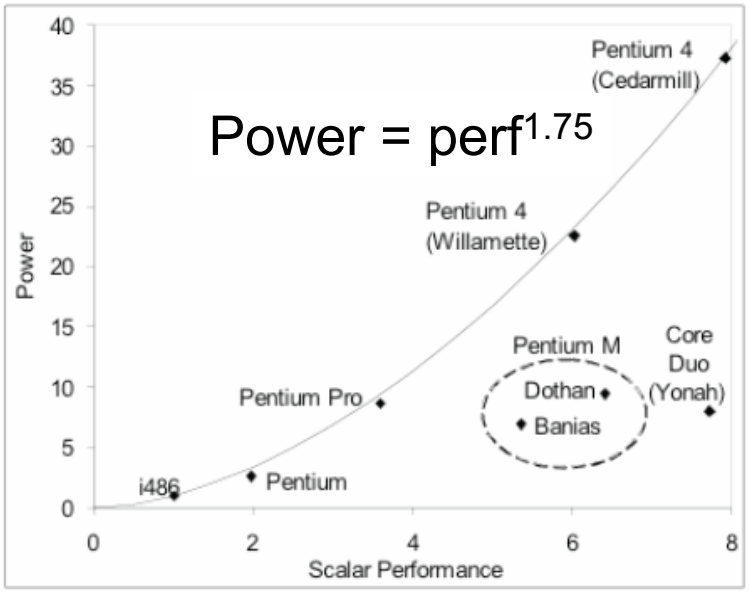


Figure 5.29: Relationship between power consumption and performance for Intel processors normalized for fabrication technology. [Grochowski_06]

DSPs have also been exhibiting a growing efficiency in terms of the operation rate that can be supported per mW by increasing the number of operations executed per cycle. As shown in Figure 5.30, fixed point processors, which are better able to apply SIMD functions. have exhibited significantly more improvement in terms of the number of 16-bit operations it can complete per second per mW. A similar effect is also seen in Figure 5.31 for 16-bit MACs. That this trend is largely due to increases in parallelism is reinforced by the picoChip PC102 (shown in yellow in Figure 5.31) which can execute a total 38.6 GMACs while its core consumes 4.375 W (5 W total – 625 mW I/O) [PC_04].[12]

[12] While there are more recent picoChip products available, the PC102 is the most recent whose documentation provides power consumption metrics. The various picoChip products also provide MIPS measurements, but these are for RISC (Reduced Instruction Set Computer) operations and generally vary greatly in terms of the number of cycles required for execution which makes comparisons with the single-cycle MAC chips surveyed difficult. The PC102 MAC numbers, however, are explicitly for single-cycle 16x16 MACs of the form considered throughout this section, thereby making direct comparisons possible.

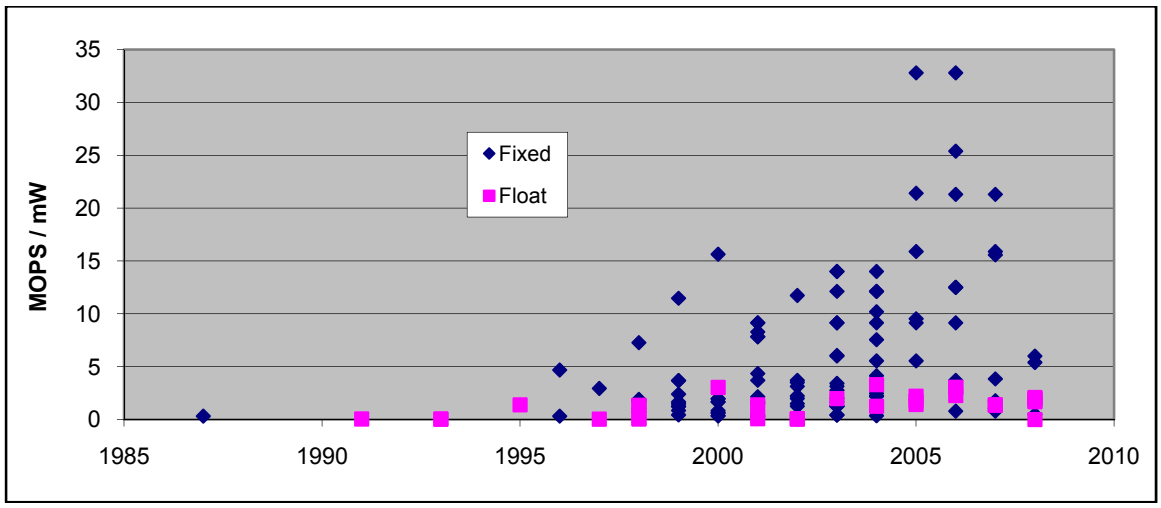


Figure 5.30: Fixed point processors have exhibited dramatic improvements in computational efficiency.

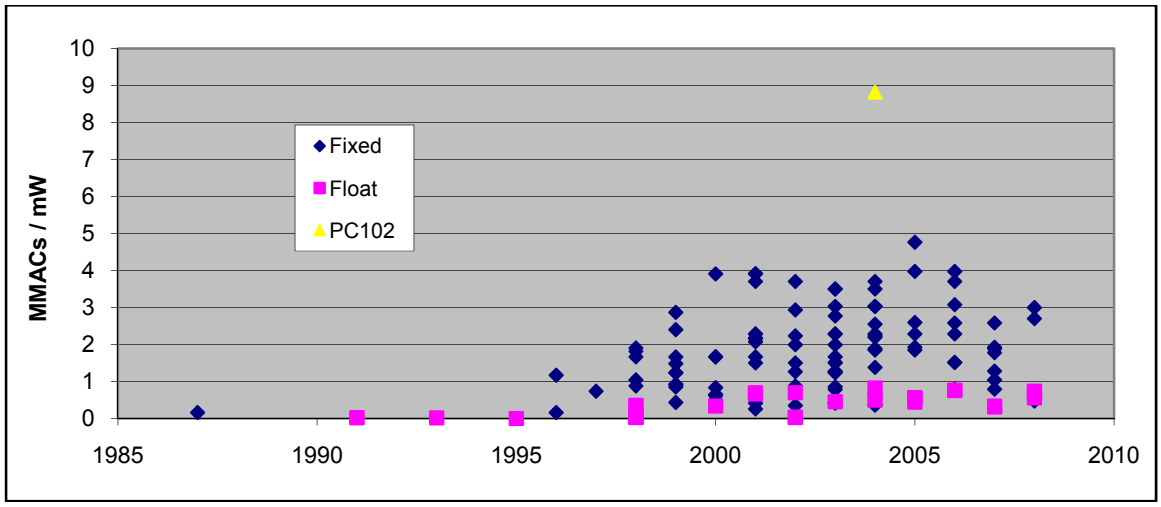


Figure 5.31: While DSPs have become noticeably more computationally by performing more operations per cycle, picoChip's PC102 massive multicore architecture is significantly more efficient.

5.2.2 Implications for SDR Design and Implementation

The following discusses the expected impact of increases in performance, the trends to increasing parallelism, and the discussed trends of transistors.

5.2.2.1 Impact of Increases in Performance

As discussed in Section 5.2.1, even though GPP clock speeds have saturated, Moore's Law is still working to the advantage of SDRs by steadily increasing both computational capacity and computational efficiency. In short, this means that SDRs can continue to support ever increasing complex waveforms without significantly increasing power consumption or potentially support the same waveforms with less power.

While the increasing parallelism will likely lead to changes in software development as discussed in Section 5.2.2.2 will slow this shift, the obvious superiority of massive parallelism in terms of computational efficiency as exhibited the PC102 in Figure 5.31 implies that this trend will ensure that processors will become massively parallel.

5.2.2.2 Impact of Trend Towards Multicore Processors

As noted in Section 5.2.1.5, both GPPs and DSPs have been trending towards the use of multiple embedded cores and more generally, more operations per cycle. As chip architectures continue to change, it is unlikely that existing software will be able to take advantage of this increasing parallelism. Generally, exploiting this parallelism will require redesigns of radio software that has traditionally been operated serially. Fortunately, some real-time-operating systems have been developed to support multi-core processors, e.g., [Enea_04] and [Ireton_06], which should ease this process. Nonetheless, the following are issues that programmers will need to be aware of in a multi-core world: memory and resource contentions, program locality[13], and inter-core timing issues (e.g., deadlock where cores are waiting on each other and race conditions).

5.2.2.3 Impact of Continued Transistor Trends

To estimate how the power, speed, and performance of key SDR components (ADCs and DSPs) will change as transistors continue to scale, we apply the traditional scaling relationships shown in Table 5.11. In full scaling (also known as constant electric field scaling), all dimensions are scaled by the same factor S (S>1); in fixed voltage scaling (fixed-V in Table 5.11) all dimensions are scaled by S except for voltages which are held constant; and in general scaling all dimensions are scaled by S except for voltages which are scaled by a factor U (U>1). For our purposes, the key parameters are the increase in circuit speeds (decrease in intrinsic delay) and the effect on power consumption.

[13] In theory, the operating system should handle this, but this will generally be less optimal than hand-coding.

Table 5.11: Key Scaling Relationships [Patel_05]

Parameter	Relation	Full Scaling	General Scaling	Fixed-V Scaling
W, L, t_{ox}		1/S	1/S	1/S
V_{DD}, V_T		1/S	1/U	1
N_{SUB}	V/W_{depl}^2	S	S^2/U	S^2
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	C_{ox} WL	1/S	1/S	1/S
k_n, k_p	C_{ox} W/L	S	S	S
I_{sat}	C_{ox} W V	1/S	1/U	1
Current Density	I_{sat} / Area	S	S^2/U	S^2
R_{on}	V / I_{sat}	1	1	1
Intrinsic Delay	$R_{on}C_{gate}$	1/S	1/S	1/S
P	$I_{sat}V$	$1/S^2$	$1/U^2$	1
Power Density	P/Area	1	S^2 / U^2	S^2

While the intrinsic delay and power rows appear to imply that speed and power can be varied independently (via the choices for *S* and *U*), this is not the case as transistor delay is given by

$$T_D \approx \frac{C_L V_{DD}}{\varepsilon \left(V_{DD} - V_t\right)^2} \qquad (0.2)$$

where C_L is the capacitance along the critical path, ε is a device-specific parameter, and V_t is the threshold. Note that T_D is effectively inversely related to V_{DD} such that as V_{DD} decreases T_D increases. Thus, scaling to decrease power consumption will tend to work against scaling to increase speed, though this tradeoff can be somewhat mitigated via improvements in fabrication technology (which changes ε).

Since we are clearly not free to arbitrarily choose values for *S* and *U*, for projection purposes it is useful to examine the choices of chip designers. As shown in Figure 5.32, scaling for the last few decades has relied on a mix of general scaling and fixed-voltage scaling. Alternately, this can be viewed as continual general scaling with $S > U \geq 1$. For our purposes, we will model transistor trends as operating under general scaling with *S* as $2^{1/4}$ each year (to accommodate the transistor doubling every two years) or *S* = 1.19 and *U* as 1/0.9 or 1.11 based on the voltage trends from 1995 to 2003 illustrated in Figure 5.32.

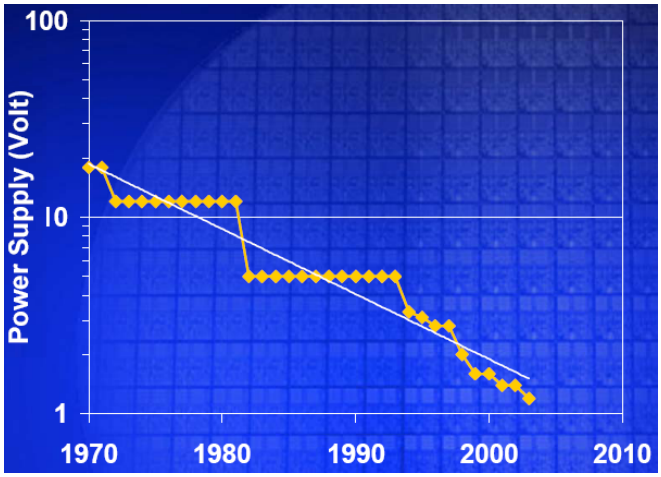


Figure 5.32: Progression of Intel Processor Supply Voltages. [Moore_03]

Using ***S* = $2^{1/4}$ or 1.19** and ***U* as 1/0.9 or 1.11** we can quantify relationships for a number of important parameters as shown in Table 5.12 where *y* represents the number of years of scaling. Note that this implicitly assumes continuing chip frequency increases. If, instead chip frequencies are held constant, chip power will instead scale as $(S/U^2)^y$ and chip power consumption will actually decrease even though the number of transistors continue to increase. Also note that *y* should not assume values greater than 16 based on the discussion in Section 5.3.1.1.

Table 5.12: Modeling parameters for projecting transistor trends

Parameter	General Scaling	Scaling Values
Transistor Speed	$(S)^y$	1.19^y
Power / transistor	$(1/U^2)^y$	0.81^y
Number transistors / Chip	$(S^2)^y$	1.414^y
Power Density	$(S^2/U^2)^y$	1.146^y

While we will use these equations, it is important to realize that they are primarily intended to model **CMOS** (**C**omplementary **M**etal **O**xide **S**emiconductors) integrated circuits and as shown in Figure 5.33, furtherance of these trends will likely require significant technology changes. For our purposes, however, we continue to use the scaling relationships of Table 5.12 as a more detailed model of unknown exotic materials seems impossible. Of greatest consequence, this means that the projected power levels presented in this section underestimate eventual power consumption levels.

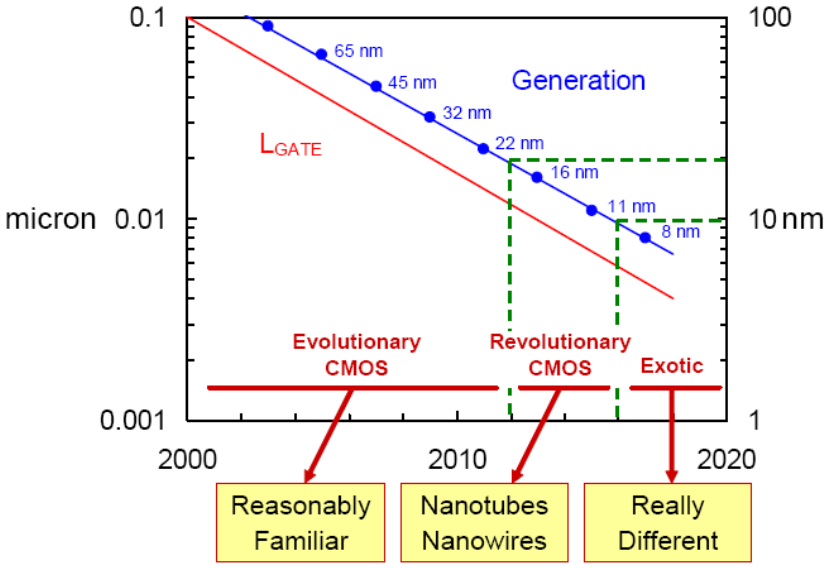


Figure 5.33: Projected Nanotechnology Eras. While Moore's law is projected to continue into the foreseeable future, technologies will become increasingly exotic. [Gargini_04]

The expected movement towards more exotic materials highlights the trends discussed in Section 5.2.1.6 where we saw that revenues were relatively flat while fabrication costs were exponentially increasing. In practice, the combination of increasing fabrication costs and flat revenues means that manufacturers will have to go to greater lengths to increase the return on investment on lithography tools. Some techniques to achieve this that could impact SDR design and implementation include the following:

- Consolidate fabrication efforts so that a single set of tools addresses many different chipsets. This solution is effectively at the heart of fab-less semiconductor companies for whom the costs of their own set of tools cannot be justified.
- To the extent that mask costs are also rising, develop chipsets that attempt to capture larger markets. In the wireless market, this implies a greater emphasis on processors suitable for SDR.
- Extend the useful life of fabrication tools. It would be expected that this trend would be realized by keeping popular chipsets on the market for longer and by "making-do" with transistor feature sizes that are not cutting-edge.

5.3 Fundamental Limits to DSP Performance

The following discusses the fabrication and physical limits to DSP performance and the implications to SDR design and implementation.

5.3.1 Sources of Fundamental Limits

As shown in Figure 5.34, there is a fundamental tradeoff between operations / second and power consumption, which depends on processor architectures and transistor technologies. While we previously noted that Moore's Law has continued to decrease feature sizes and to increase the number of transistors per device (Section 5.2.1.4), there will be a limit to how small transistors can become before the physics breaks down. Such a limit does not mean that devices built on a different set of physical relationships (e.g., quantum computing) will not emerge later and change this limit, but extrapolations beyond this limit would have little empirical basis at this time. The following discusses the role these limits play in processor performance.

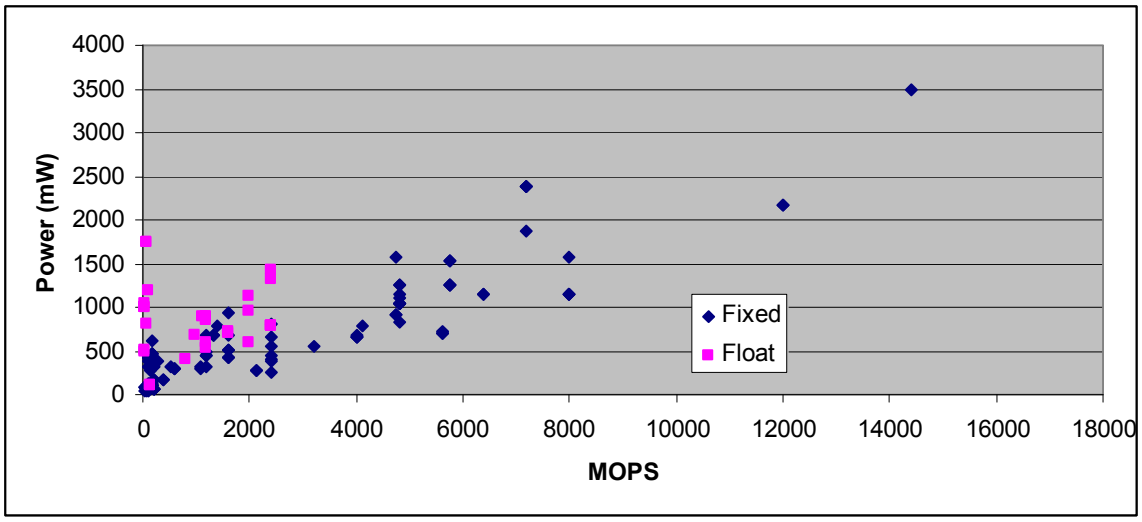


Figure 5.34: Power is strongly correlated with computational capacity.

5.3.1.1 Estimated End of Moore's Law

For decades integrated circuit transistor density has been approximately doubling every two years (see Figure 5.35).[14] This scaling phenomenon, and the associated trends in speed and price, dramatically reshaped the world as existing and many never-before-possible functionalities were shifted to silicon. This led to a phenomenon where device performance improved as costs declined.

[14] Integrated circuits are largely two-dimensional structures. This means that to increase density by a factor of 2, feature size must be scaled by a factor of about 0.7. Feature size then decreases by a factor of 2 every 4 years.

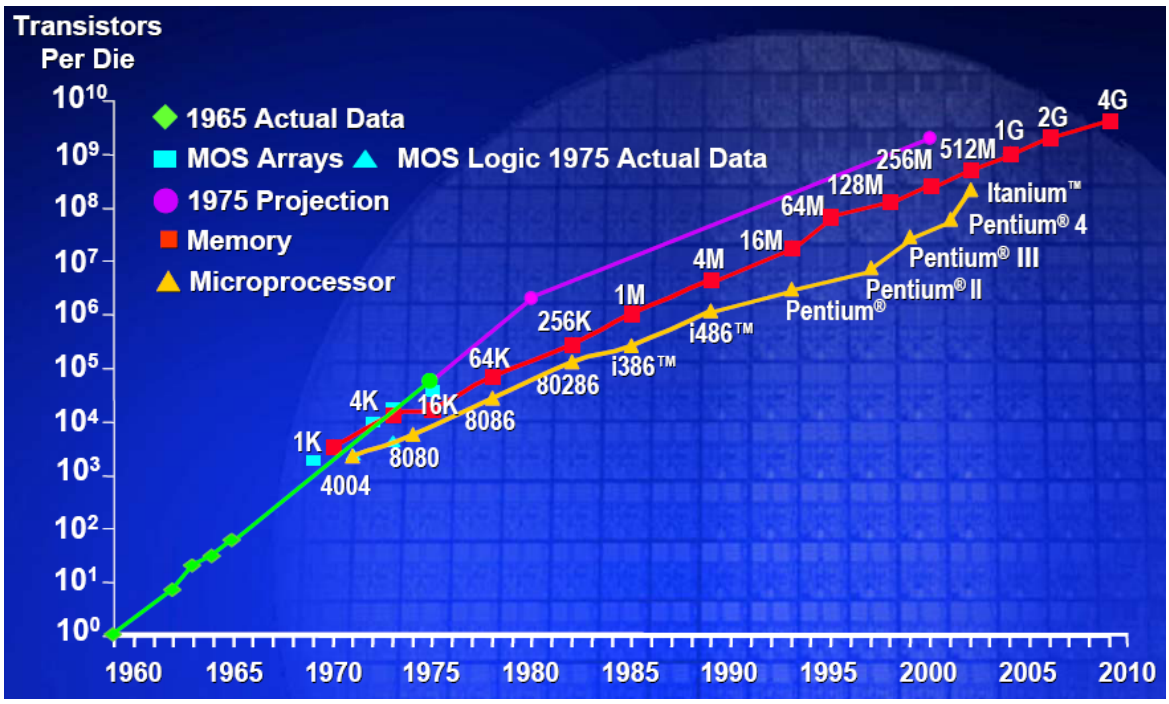


Figure 5.35: Across various platforms, transistor density has approximately doubled every two years. [Moore_03]

For almost as long as Moore's law has been in effect, prognosticators have predicted its imminent demise due to some potentially insurmountable technical hurdle. For example, Dr. Suman Datta recently projected that only four years remained for Moore's Law with the caveat that other technologies (e.g., superconductors and carbon-nanotubes) might change this projection [Gardiner_08]. While this pronouncement received significant hype, in reality this is the normal condition of the semiconductor industry – Moore's Law has consistently needed radical improvements in technologies to progress beyond the next two or three generations.

For example, the shift to High-k insulators and metal gates was a previously unproven approach. While significant modifications to manufacturing processes were necessary to make these changes, both Intel (45 nm) and NEC (55 nm) were able to successfully include these in their manufacturing processes and shipped products with Hafnium gate insulators and metal gates in 2007 [West_07] [Leopold_07]. Measurements of these products have shown that this shift reduced leakage current by a factor of 100 [Gargini_08]. Looking backwards, this transition in 2007 was predicted by Intel in 2003 as summarized Figure 5.36. Thus, even though the manufacturing precise processes were not known at the time, Intel's experience of continually altering and improving their fabrication process led to their achieving their process goals right on schedule.

Process Name	P856	P858	Px60	P1262	P1264	P1266	P1268	P1270
1st Production	1997	1999	2001	2003	2005	2007	2009	2011
Process Generation	0.25µm	0.18µm	0.13µm	90 nm	65 nm	45 nm	32 nm	22 nm
Wafer Size (mm)	200	200	200/300	300	300	300	300	300
Inter-connect	Al	Al	Cu	Cu	Cu	Cu	Cu	?
Channel	Si	Si	Si	Strained Si	Strained Si	Strained Si	Strained Si	Strained Si
Gate dielectric	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	High-k	High-k	High-k
Gate electrode	Poly-silicon	Poly-silicon	Poly-silicon	Poly-silicon	Poly-silicon	Metal	Metal	Metal

Figure 5.36: Intel Process Projections from 2003. [Gargini_08]

Other Upcoming Technologies

While the projection in Figure 5.36 only goes out to 2011, numerous technologies for use in subsequent features have already been identified as shown in Figure 5.37. These include spin devices, optical wires, molecular electronics, carbon and nanotubes. Because of past history and the numerous technologies in the pipeline, we feel confident that solutions will continue until more fundamental limits are reached.

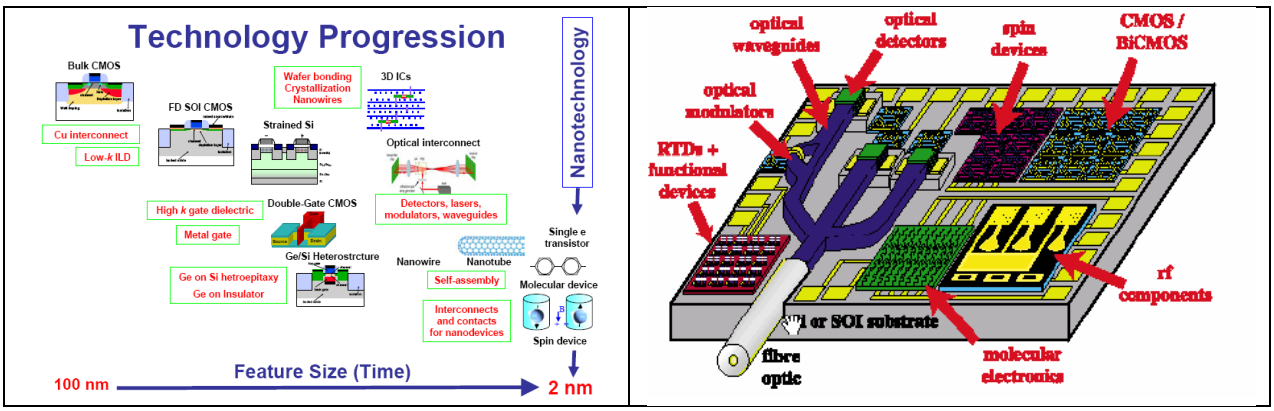


Figure 5.37: Numerous different technologies have been identified as candidates for use as feature sizes continue to decrease. [Gargini_04]

A more concrete assessment in light of these technologies is given by the **I**nternational **T**echnology **R**oadmap for **S**emiconductors (**ITRS**) which annually conducts industry-wide surveys to identify key technical challenges and project the state of the transistor industry 15-years out. As summarized in Figure 5.38, while significant technical hurdles have been identified, the ITRS prognosticators believe transistor scaling will continue until at least 2022.

Table B ITRS Table Structure—Key Lithography-related Characteristics by Product

Near-term Years

YEAR OF PRODUCTION	2007	2008	2009	2010	2011	2012	2013	2014	2015
DRAM stagger-contacted Metal 1 (M1) ½ Pitch (nm)	65	57	50	45	40	36	32	28	25
MPU/ASIC stagger-contacted Metal 1 (M1)½ Pitch (nm)	68	59	52	45	40	36	32	28	25
Flash Uncontacted Poly Si ½ Pitch (nm)	54	45	40	36	32	28	25	23	20
MPU Printed Gate Length (nm)	42	38	34	30	27	24	21	19	17
MPU Physical Gate Length (nm)	25	23	20	18	16	14	13	11	10

Long-term Years

YEAR OF PRODUCTION	2016	2017	2018	2019	2020	2021	2022
DRAM stagger-contacted Metal 1 (M1) ½ Pitch (nm)	22	20	18	16	14	13	11
MPU/ASIC stagger-contacted Metal 1 (M1) ½ Pitch (nm)	22	20	18	16	14	13	11
Flash Uncontacted Poly Si ½ Pitch (nm)	18	16	14	13	11	10	9
MPU Printed Gate Length (nm)	15	13	12	11	9	8.4	7.5
MPU Physical Gate Length (nm)	9	8	7	6.3	5.6	5.0	4.5

Figure 5.38: Summary of ITRS Projections. [ITRS_07]

A More Fundamental Limit?

While all previous predicted ends to Moore's law due to an "insurmountable" obstacle have been overcome to date, there may be a more fundamental limit to the transistor scaling process, a limit endorsed by Gordon Moore. In 2007, Gordon Moore estimated that Moore's law would only last another 10 to 15 years [Gardiner_07] which would mean that the ITRS projections of Figure 5.38 would be overly optimistic. To support this judgment Moore cited Stephen Hawking's suggestion that the size of an atom and the speed of light would be fundamental limits to transistor performance. These same limiting factors were presented in [Gargini_04] which presented the model shown in Figure 5.39 for what could be the smallest possible transistor – effectively 5 molecules across (insulator, source, channel, drain, and insulator) and gave an estimated molecular size of 5.4 Angstroms. Assuming the gap between transistors is the width of a single insulator molecule, each transistor would be four molecules wide (by sharing insulators) which would give a **fundamental minimum transistor width of 2.1 nm** (4 x 5.4 Angstroms).

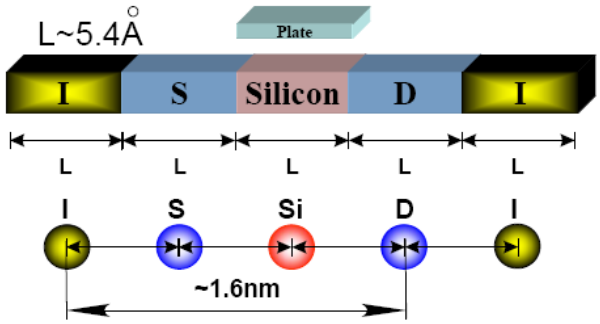


Figure 5.39: Assuming transistors scaling continues to the placement of individual molecules, minimum transistor size will be approximately 2.1 nm (4xL). [Gargini_04]

Using Intel's current (2007) 45 nm generation as a reference and assuming the continued progression of dimension scaling of a factor of 0.7 every two years, 8 more scalings would be possible before transistors would fall below this fundamental limit. **This yields a projected end to Moore's Law in 16 years or 2023**, which is just outside of Moore's estimate but in line with current ITRS projections.

Remaining Gains in Performance

Based on the preceding discussion, we estimate that without a fundamental shift in transistor and processor physics (e.g., quantum computing) Moore's law will come to an end in 16 years. If we substitute that time period into the expressions of Table 5.12 and assume the same scaling relationships of the past few years continue, we can solve for the estimated remaining gains in transistor performance shown in Table 5.13 where we estimate transistor speeds to increase by another factor of 16, transistor density to increase by a factor of 256, power / transistor to only be 3.4% as much as today (which ignores possible further rises in leakage current), and an increase in power density by a factor of 8.79. Note that the already high power density of GPPs means that this increase would likely not be possible without significant improvements in heat transfer technology which implies that for GPPs, at least, there will also be a fundamental shift in scaling factors (perhaps a greater emphasis on U instead of S, i.e., more emphasis on reducing voltages and less on speed). This also means that, at least for GPPs, a further speed increase by a factor of 16 is overly optimistic.

Table 5.13: Estimated limits to further gains in transistor performance.

Parameter	General Scaling	Scaling Values	Remaining Gains
Transistor Speed	$(S)^y$	1.19^y	16 x
Power / transistor	$(1/U^2)^y$	0.81^y	3.4%
Number transistors / Chip	$(S^2)^y$	1.414^y	256 x
Chip Power Density	$(S^2/U^2)^y$	1.146^y	8.79 x

5.3.1.2 Computational Efficiency Limits

Power consumption values for processors include many important activities that are nonetheless not directly in service of computation, e.g., chip I/O, the instruction fetch / decode / dispatch operations, memory accesses, and peripherals. So to compute a lower bound to computational efficiency, let us assume that a processor consists of only a multiplier and an ALU. Again as with processors, there are an infinite number of ways to architect a multiplier and an ALU and the optimal architecture is strongly dependent on the application. Nonetheless, by excluding so

much of a processor's components that are not directly applied to computation[15], any choice of multiplier and ALU should serve as a lower bound.

Based on the preceding, we compute a lower bound for power per MACS from the total power consumed by a multiplier and ALU. For a multiplier, we turn to [Soni_01] which estimated the implementation of a signed 16x16 multiplier as consuming 10.9 mW with a supply of 3.3 V and clocked at 50 MHz, with a minimum feature size of 250 nm. For an ALU, we turn to [Matthew_04] which estimated the implementation of a 32-bit ALU as consuming 75.4 mW with a supply of 1.3 V and clocked at 7 GHz, with a minimum feature size (fabrication technology) of 90 nm.

To eliminate the effects of using different fabrication technologies, different supply voltages, and different clocks, we normalize the multiplier's parameters to the ALU parameters as follows. [16] First, referring back to (0.30), we scale the multiplier power consumption estimate by a factor of (7000 / 50) x $(1.3/3)^2$ = 26.3 for 26.3 x 10.9 mW = 288 mW. Then to estimate the gain in terms of ε, using a value for $U = 0.9$, the effect on power consumption due to changes in fabrication technology from 250 nm to 90 nm is equivalent to 6 years of advancement according to Figure 5.36 (1997 – 2003) means that power per transistor would decrease by a factor of 0.9^6 = 0.53. This gives an estimated power consumption value for the multiplier reported in [Soni_01] if run at the clock rate and voltage level and using the fabrication technology of [Matthew_04] of 153 mW.[17]

Thus at a clock rate of 7 GHz, this MAC-only chip would consume (153 + 75.4) = 228.4 mW for a MMAC / mW ratio of 7000 MMACs / 228.4 mW= 30.64. If we again make use of the values for S and U discussed in Table 5.12, where speed increases as S^y and power / transistor scales as $(1/U^2)^y$, then we can derive an expression for a scaling factor to bound the MMAC / mW ratio as a function of fabrication technology as $(S \times U^2)$, or an improvement by a factor of 1.46 every year. Since the initial bound was for 90 nm features (4 years ago) we would expect the performance ratio bound to saturate at (1.46^20) x 30.64 = **66,334 MMACS / mW**. Equivalently, and perhaps more intuitively, we can also express this ratio in terms of mW / MMAC by evaluating the inverse to yield approximately **15 nW for 1 MMACS**.

Based on this analysis, we can plot the curve defined by changes in fabrication technology as shown in Figure 5.40 which also includes a scattering of estimated values for surveyed DSPs that reported core power levels, MMACS, and fabrication technology (primarily Texas Instruments parts as the other vendors rarely reported the fabrication technology). Note that the scatter plot

[15] This is another factor in the relative computational efficiency of FPGAs viz a viz DSPs – a greater faction of an FPGA's transistors are directly devoted to computation. An ASIC, of course, is more efficient still as an even greater fraction of its transistors are applied to computation.

[16] The use of 16-bit multipliers with 32-bit ALUs is not unusual. For example see Texas Instruments' C62xx line of processors.

[17] This is a dramatic oversimplification ignoring second-order effects. However, as the only attempt here is to give an upper bound to MMACS / mW, these implicitly upward biasing errors only serve make the bound less tight.

(except for the PC102 outlier) loosely follows the same curve as the 1 multiplier, 1 ALU hypothetical DSP considered, but necessarily at a higher level due to hypothetical DSP's exclusion of instruction and memory handling. Also as expected, fixed-point processors are generally seen to be more efficient than floating point processors when fabrication technology is held constant.

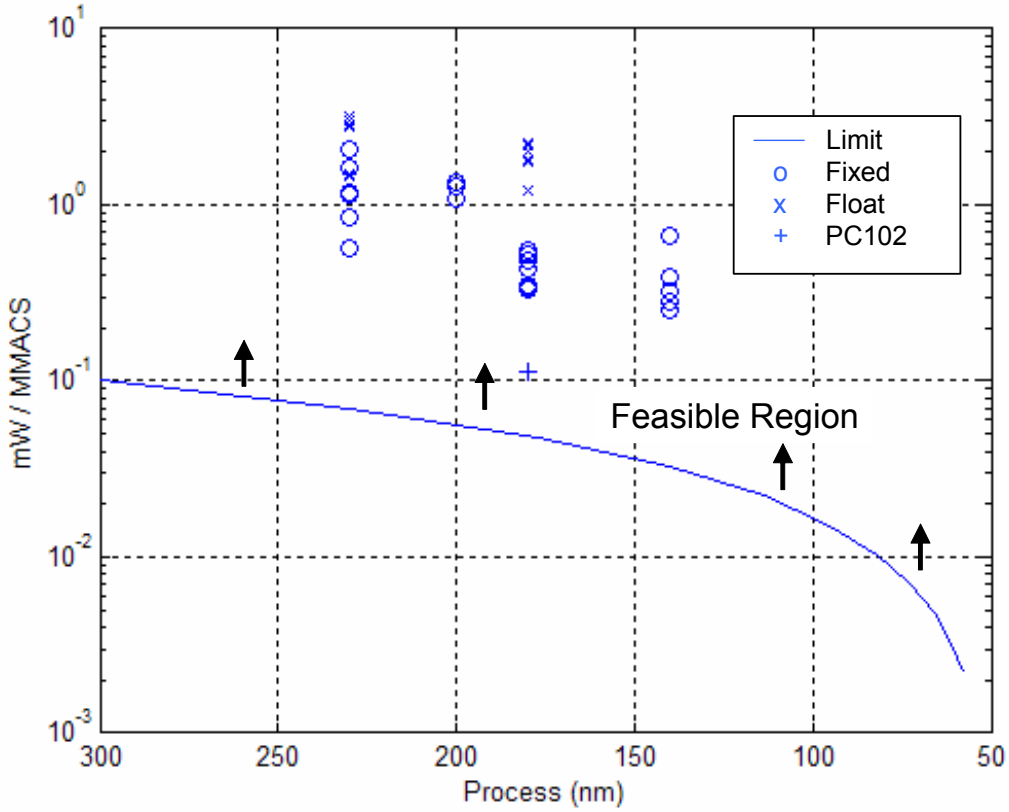


Figure 5.40: The limit to the amount of power required per MMACS is a function of fabrication technology.

5.3.2 Implications for SDR Design and Implementation

Without a major change in the physics of transistors, transistor size will eventually be limited to the width of a handful of atoms. This means that we should not expect Moore's Law to make every algorithm feasible on a single chip of today's size and power. Further, based on the projections shown in Section 5.3.1.1, if an algorithm would require operation at 16x faster than the fastest chip currently available, then the algorithm's eventual feasibility along the natural progression should be doubted. This will likely be a significant hurdle in the years to come.

Without advances in Moore's Law to leverage, further performance gains will have to come from changes in processor architectures and changes in the way algorithms are deployed. Already, we have seen a significant shift to multi-core processors and more generally to architectures that execute numerous operations per cycle to avoid power density issues. There is

some hope for significant gains in computational efficiency due to changes in processor architecture because of the gap in performance per unit power shown in Figure 5.40 and the outlier effect of the PC102. Further gains may also be possible via the thin-client model, but as applied to radios to allow major computations (e.g., policy reasoning) to occur in an external device with greater computational and cooling capacity. A third possible avenue to circumvent these limits would be for the development of room-temperature super-conductors suitable for use inside of DSPs. [18]

As architectures grow increasingly parallel and are optimized for different applications, it seems likely that it will be necessary to develop more flexible programming styles and practices. This should not be an insurmountable hurdle as it can be analogized to the previous migration from procedural to object-oriented code, but it will take some time for practices to adjust to the new realities. Alternately, the necessary parallelizations and optimizations could be performed in more advanced compilers to allow programming to continue at a high-level while targeting (and optimizing for) a changing set of highly-parallel processor architectures. Such a solution would also be valuable in addressing current code portability issues in SDRs. However, based on current DSP compilers intended for superscalar and SIMD architectures, there will likely be a large gap between hand-coded and compiled performance which will mean that the most demanding implementations will be the least portable.

REFERENCES

[Altera_03] Altera, "An Analytical Review of FPGA Logic Efficiency in Stratix, Virtex-II, & Virtex-II Pro Devices," *Altera White Paper*, May 2003, Available online: http://www.altera.com/literature/wp/wp_stx_logic_efficiency.pdf
[Altera_05] Altera, "Stratix II and Virtex IV Power Comparison & Estimation Accuracy," *Altera White Paper*, August, 2005, Available online: http://www.altera.com/literature/wp/wp_s2v4_pwr_acc.pdf
[BDTI_06] "BDTI Pocket Guide," BDTI, http://www.bdti.com/pocket/pocket.htm
[Brodkin_08] J. Brodkin, "Shift to multicore processors inevitable, but enterprises face challenges," *LinuxWorld*, February 2008. Available online: http://www.linuxworld.com/news/2008/022708-multicore-processors.html
[Commstack_03] "Why License Commstacks OFDM Modem IP Cores," CommStack white paper, 2003. Available online: http://www.commstack.com/WhyLicenseCommStacksOFDMModemIPCoresv2.pdf
[Crocket_98] J. Crocket, "DSP Architectures for Wireless Communications," 1998 International Symposium on Advanced Radio Technologies.
[Enea_04] "Enea Supports TI's OMAP with Bundled Multi-Core RTOS Platform," ENEA Press Release, March 31, 2004. Available online: http://www.embeddedstar.com/press/content/2004/3/embedded13813.html
[Gardiner_07] B. Gardiner, "IDF: Gordon Moore Predicts End of Moore's Law (Again)" Wired, Sep 2007, http://blog.wired.com/business/2007/09/idf-gordon-mo-1.html
[Gardiner_08] B. Gardiner, "Tick-Tock: Researcher Says Silicon Chips Have Four Years of Improvement Left," *Wired*, March 2008, Available online: http://blog.wired.com/business/2008/03/researcher-says.html
[Gargini_04] P. Gargini, "Sailing with the ITRS into Nanotechnology," *Semicon West*, July 12, 2004.
[Gargini_08] P. Gargini, "Overcoming the Red Brick Walls," *Industry Strategy Symposium*, Jan 2008. Available online: http://download.intel.com/technology/silicon/orbw.pdf
[Grochowski_06] E. Grochowski and M. Annavaram, "Energy per Instruction Trends in Intel Microprocessors," *Technology @ Intel Magazine*, March 2006. Available online: ftp://download.intel.com/pressroom/kits/press/core2/epi-trends-final2.pdf
[Ireton_06] M. Ireton and M. Kardonik, "Using a multicore RTOS for DSP applications," DSPDesignLine, July 17, 2006. Available online: http://www.embedded.com/columns/showArticle.jhtml?articleID=190500287
[ITRS_07] "Executive Summary," The International Technology Roadmap for Semiconductors: 2007, Available online: http://www.itrs.net/Links/2007ITRS/ExecSum2007.pdf

[18] Actually, they would have to be superconducting well-above room temperature for operation in a normal DSP.

[Leopold_07] G. Leopold et al., "High-k push brings anxiety as 32 nm looms," EE Times, December 17, 2007. Available online: http://www.eetimes.com/news/semi/showArticle.jhtml?articleID=204803606
[Linux_07] "The Core of the Issue: Multicore and You," *Linux Magazine*, November 13, 2007. Available online: http://www.linux-mag.com/id/4311
[Matthew_04] S. Matthew, et al., A 4GHz 300mW 64b Integer Execution ALU with Dual Supply Voltages in 90nm CMOS," *ISSCC 04*.
[Maxfield_08] C. Maxfield, "Xilinx Responds to Altera's Benchmarks," *Programmable Logic Design Line*, May 30, 2008. Available online: http://www.pldesignline.com/blogs/208401153
[McKenney_07] P. McKenney, "SMP and Embedded Real Time," *Linux Journal*, January 1, 2007. Available online: http://www.linuxjournal.com/article/9361
[Moore_03] G. Moore, "No Exponential is Forever," International Solid State Circuits Conference 2003, Available online: http://download.intel.com/research/silicon/Gordon_Moore_ISSCC_021003.pdf
[Neel_02] J. Neel, "Simulation of an Implementation and Evaluation of the Layered Radio Architecture," MS Thesis, Virginia Tech December 2002.
[Neel_04] J. Neel, S. Srikanteswara, J. Reed, P. Athanas, "A Comparative Study of the Suitability of a Custom Computing Machine and a VLIW DSP for Use in 3G Applications," *IEEE Workshop on Signal Processing Systems SiPS2004,* Oct 13-15, 2004, pp. 188-193.
[Neel_05] J. Neel, P. Robert, J. Reed, "A formal methodology for estimating the feasible Processor solution space for a software radio", *SDR Forum Technical Conference 2005*, Orange County, CA, Nov. 14-18, 2005, #1.2-03.
[Patel_05] C. Patel, "Lecture 11: Scaling," UMBC, October 2005. Available online: http://www.csee.umbc.edu/~cpatel2/links/640/lectures/lect11_scaling.pdf
[PC_04] "PC102 Product Brief," PicoChip, March 2004.
[Rivoallon _02] F. Rivoallon, "Comparing Virtex-II and Stratix Logic Utilization," Xilinx White Paper 161, June 2002, Available online: http://www.xilinx.com/support/documentation/white_papers/wp161.pdf
[SDRF_08] SDR Forum "Cognitive Radio Definitions and Nomenclature," Working Document SDRF-06-P-0009-V0.5.0, May 30, 2008.
[Silva_06] M. Silva, J. Ferreira, "Support for partial run-time reconfiguration of platform FPGAs," *Journal of Systems Architecture: the EUROMICRO Journal*, vol 52 issue 12, pp. 709-726, December 2006.
[Soni_01] M. Soni, "VLSI Implementation of a Wormhole Run-time Reconfigurable Processor." MS Thesis, Virginia Tech, 2001.
[TI_07] "Texas Instruments Goes Beyond 3G with New LTE Offering for Wireless Infrastructure," *Texas Instruments Press Release*, April 2007, Available online: http://focus.ti.com/pr/docs/preldetail.tsp?sectionId=594&prelId=sc07077
[TI_08] Texas Instruments, "TMS320C6414T, TMS320C6414T, TMS320C6414T Fixed-Point Signal Digital Signal Processors," *Texas Instruments Datasheet SPRS226L*, February 2008.
[West_07] J. West, "Intel Ships New Hafnium Assisted Transistors," Inside HPC, November 2007, http://insidehpc.com/2007/11/12/intel-ships-new-hafnium-assisted-transistors/
[Wiki_08] "Transistor Count," Wikipedia, http://en.wikipedia.org/wiki/Transistor_count
[Xilinx_07] Xilinx, "Virtex-II Platform FPGAs: Complete Data Sheet," DS031 (v3.5), November 2007, Available online: http://www.xilinx.com/support/documentation/data_sheets/ds031.pdf

6 SDR Execution Latency

6.1 Latency in Software Defined Radios

The software radio concept is built upon the use of reconfigurable (programmable) hardware whose operation can be changed through software modifications. Traditional computing devices include General Purpose Processors (GPP), Application Specific Integrated Circuits (ASIC), Field Programmable Gate Arrays (FPGA), and Digital Signal Processors (DSP). ASICs are highly optimized for specific applications in power, computing, and size, but they are lacking in flexibility and reconfiguration ability. FPGAs are power efficient and to some extent reconfigurable while very hard to program for people who don't know hardware language well. Communications DSP chips for signal processing have good flexibility although high performance DSP chips are expensive. Current GPPs are fast enough to do a lot of real time digital signal processing tasks and functions. With many library functions and a very friendly development environment, GPPs appear in several widely used SDR architectures, like GNU Radio [1], OSSIE [2], Software Communication Architecture (SCA) [3], and Space Telecommunications Radio System (STRS) [4], and can achieve very good reconfigurability.

In selecting computing hardware for SDR development, a designer must trade off factors like power efficiency, computation capacity, reconfiguration, development environment, size, and price. GPPs have the advantage of an easier development environment and faster reconfiguration than any other platforms, therefore offering the potential of real-time multi-band multi-mode reconfiguration in SDR. However, adopting GPPs in SDR development change the design process from being just a radio or DSP problem anymore. It is also a computer problem; GPP's architecture [5] and operating system (OS) mechanism [6] should be considered. ASICs, FPGAs, and DSPs don't have such issues. For example, the memory hierarchy in GPPs, as shown in Figure 6.1, introduces latency uncertainty. Furthermore, once reconfiguration happens, SDR execution latency can be a particularly troublesome issue.

For ASICs, FPGAs, and DSPs, latency is primarily related to computing capacity which is easy to quantify for radio function's computational requirement. Latency is also affected by many other factors in GPPs. GPPs have a memory hierarchy as shown in Figure 6.1. GPPs run any program instructions (with data) in the CPU with registers while instructions and data are usually stored in the disk or memory. There is a vast speed difference between CPU and memory [5]. To solve this problem, several levels of caches are inserted between the CPU and memory, and speculative methods are used to pre-fetch instructions or data into the caches. However, any failure in speculative pre-fetching will cause the CPU to wait for data read from memory with a long delay, as shown in Figure 6.1.

Run-time uncertainty also complicates SDR design [7]. Specifically, modern GPP OS is designed to maximize resource utilization – to assure that all available CPU time, memory, and I/O are used efficiently, and that no individual user takes more than his or her fair share [6]. To maximize resource utilization, the OS uses processes and threads to run multiple jobs simultaneously: applications, system programs, drivers, etc. The CPU scheduler manages all threads and processes according to CPU-scheduling algorithms like First-Come, First-Served (FCFS) scheduling, or Round-Robin (RR) scheduling. All scheduling algorithms balance multiple criteria like CPU utilization, throughput, waiting time, response time. The implication for SDR is that the OS will create a significant but uncertain amount of latency.

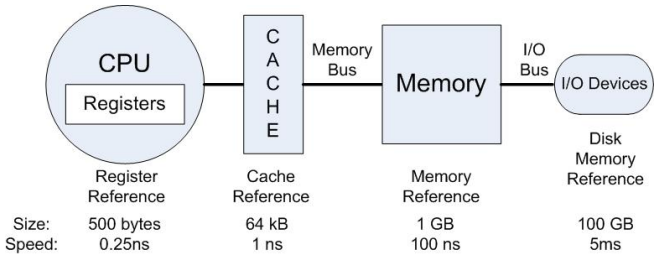


Figure 6.1: Memory Hierarchy in GPPs and the Speed Difference [5].

Furthermore, an SDR built from off-the-shelf products may have severe latency caused by I/O devices. For example, the amount of time required to send packets through a GNU Radio/USRP transmitter-receiver chain as a function of packet size and bit rate is shown in Figure 6.2 [7]. The total latency can reach the order of hundreds of milliseconds for a 1000 byte packet. Among the latency sources, the USB I/O device contributes a significant portion. Such latency will disable an SDR in supporting modern wireless standards, e.g., IEEE 802.11 and WiMAX which have very strict MAC timing requirements. Certainly there will also exist critical timing requirements for SDR functionalities that are used in cognitive radio [8] and dynamic spectrum access (DSA) developments [9].

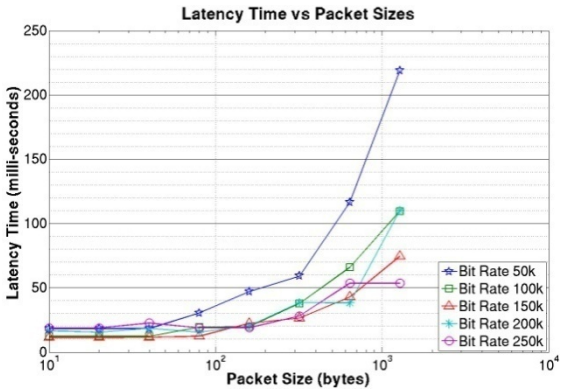


Figure 6.2: The Latency Time between Transmitting a Packet and Receiving it Using BPSK as a Function of Packet Size and Bit Rate. [7].

6.2 A Fundamental Latency Source: Pipeline vs. Sequential

Rather than analyzing the well-studied latency resulting from I/O devices, GPP memory hierarchy, and GPP OS overhead [7, 10], here we focus on a fundamental problem, that is, whether signal processing is in sequential order or in parallel/pipeline. The radios' analog part, ASICs, and FPGA all execute signals in parallel/pipeline (a continuous sequence of signals is executed simultaneously by a sequential set of components). ASICs usually run multiple function components together, and thus are in a pipeline mode. FPGAs have the advantage of dividing different slices into different functions, therefore, executing signal processing functions in a pipeline mode. But, fundamentally ,GPPs can only run one task at a time [6], even though some instruction and data level parallelisms like *instruction pipeline* and *super-scalar instruction execution* are widely implemented [5]. As a hypothetical example illustrated in Figure 6.3, the speed difference between pipeline and sequential is usually significant.

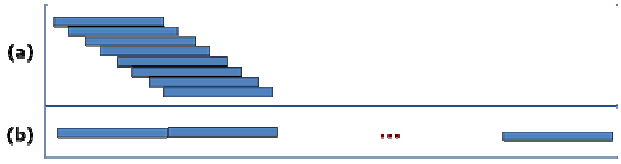


Figure 6.3: A hypothetical illustration of speed difference: pipeline (a) is 8 times faster than sequential (b).

Conventional radios (analog or digital) use a set of analog and digital components for specific radio tasks. Following an entire receiver signal processing chain, signals can continuously flow through all the analog components with negligible delay for signal processing; all the analog components process signals in a pipeline way. Usually ADC converters need very limited sampling time to get digital samples (ADCs can easily work at the order of hundreds of MHz). All of the remaining signal processing such as filtering, modulation and demodulation, source coding, and channel coding is done in the digital domain. For narrow band signals, the computational requirement is not tight; therefore, both DSPs and FPGAs can be used. For example, most public safety radios only use DSPs.

Nevertheless, it is usually very challenging for a single DSP or FPGA to support wideband waveforms, even for baseband signals. Commercial radio chips (ASICs) combine multiple computing components (system-on-chip) so that they can process multiple functions at the same time (parallel/pipeline). Certainly such a design can also get rid of latency in moving signals between different components. With an ASIC, the overall latency in the digital domain is therefore dramatically reduced. As an example, a typical Wi-Fi card [11], as shown in Figure 6.4, has several signal processing components for baseband PHY layer functions. They are used for filtering, modulation, demodulation, MAC functions, encryption, and decryption. This architecture essentially executes signal processing in a pipeline mode.

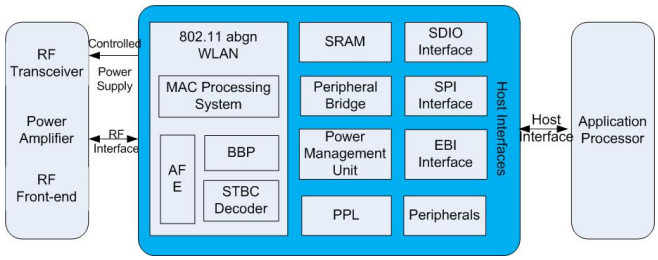


Figure 6.4: A WiFi Chip Architecture Example.

Such architecture guarantees very strict timing requirements, which is critical to IEEE 802.11's success. More specifically, there are several functions like TDMA (sync), CSMA (DIFS, SIFS), carrier sense, dependent packets (ACKs, RTS), fine-grained radio control (frequency hopping), etc., as shown in Table 6.1 that require precise and very fast timing performance. A WiFi chip must finish all the PHY/MAC layer functions within a few microseconds at both the transmitter and receiver side as well as doing all of the analog RF signal processing.

As a GPP based SDR example, GNU Radio executes all the PHY/MAC layer functions in a sequential way [7]. As we compare Figure 6.2 and Table 6.1, we can see there are two to three orders of magnitude speed difference between GNU Radio and a WiFi chip even though GNU Radio only executes narrow band signals while the WiFi chip works on wideband waveforms. As shown in [7], considering even only baseband modem functions, sequential execution of PHY layer functions takes on the order of tens of micro-seconds for a small packet. We believe the most fundamental source for such a performance gap is parallel/pipeline vs. sequential signal processing even though the GPP's architecture does introduce some additional latency overhead.

Table 6.1: Summary of important timing constants in 802.11b, 802.11a, and 802.11g [12]. (Please notice that the timing is based on performance of WiFi ASICs.)

Parameter	Value			
	802.11b	802.11a	802.11g only	802.11 g + legacy
SLOT	20 µs	9 µs	9 µs	20 µs
SIFS	10 µs	16 µs	10 µs	10 µs
DIFS	50 µs	34 µs	28 µs	50 µs
Physical Layer Header Length	192 µs [long] 96 µs [short]	20 µs	20 µs	20 µs

6.3 Implications for SDR design

In Section 6.2, we used IEEE 802.11 as an example to illustrate timing requirements in wireless standards. Since execution latency is decided by the size of computational requirement and the available computing capacity, we survey existing wideband waveforms' data rates and summarize their computational requirements for baseband signals.

The complexity of algorithms used in telecommunications to reduce the bit error rate and increase spectrum efficiency has increased continuously. For example, multi-input multi-output (MIMO) and broadband techniques have been developed to efficiently utilize both radio spectrum and energy while supporting high data rates. More channel estimation and adaption algorithms are used to reduce inter symbol interface (ISI) problem. As shown in [13], many wireless standards have already been and will continue to be created in the near future. Table 6.2 shows some wireless standards, their bit rates, and computational complexity at the baseband. Usually, higher data rates demand higher computing capacity for a same waveform. Algorithms used in smart antenna and traditional coding schemes certainly increase the computational complexity for SDR development.

Table 6.2: Wireless Standards.

Wireless Standards	Bit Rate	Computational complexity (MIPS)
GSM [14]	270.833 kbit/s	100
802.11 a [15]	54.0 Mbps	5000
CDMA2000 [16]	1.28 Mbps	2000
WCDMA [15]	3.84 Mbps	3000
TD-SCDMA [15]	1.28 Mbps	3000
OFDM-VBLAST with 4x4 MIMO (WiMAX or 802.11n) [17]	216 Mbps	9600

As Table 6.2 shows, the high data rate in wideband waveforms will not only require high bandwidth I/O devices in moving signal samples, but also demand much higher computing capacity in executing individual PHY/MAC functions. Latency performance is critical from both perspectives.

6.4 Proposed Solutions

As we can see from the above sections, SDR architecture has to achieve a similar ability in reconfiguring as GPPs have and a similar execution speed as highly optimized ASICs have. Therefore SDR should better utilize available computing resources, for example, increasing more parallelism especially considering the increasing computational requirement from future broadband wireless technologies. This is far more important than using the increasing computing ability, such as from DSP and GPPs as shown in Section 3. Therefore, we recommend that future SDR architectures proceed in two directions: hybrid architecture as shown in Figure 6.5, or a multi-core based parallel architecture.

Hybrid architecture with a control processor may contain an embedded GPP, a reconfigurable FPGA, and some auxiliary ASICs. This arrangement has several advantages for implementing an SDR. The control processor is both necessary to handle non-DSP functions such as branch, control, and decisions and efficient enough to coordinate different computing tasks and. ASICs are used for widely accepted wireless standards like WiFi, WiMAX, and LTE. Computation accelerators are for computationally intensive tasks such as video graphic processors. The FPGA can support most other PHY/MAC layer functions. Most importantly, some FPGAs support run-time reconfiguration by using techniques like Wires on Demand [18]. Therefore, upon reconfiguration request by the GPP, the FPGA is able to reconfigure very quickly by using existing function bit streams, similar to the way that library functions are used for GPP programs. Such architecture is able to reduce the power budget by using low clock frequencies, thus saving die area and reducing static and dynamic power consumption. It can achieve parallelism not only at the data and instruction level, but also at the task level by spreading tasks out among different computing components and different FPGA slices.

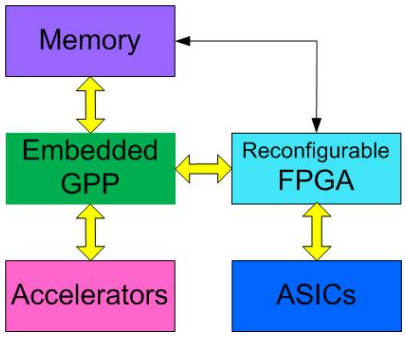


Figure 6.4: The proposed embedded GPP/FPGA hybrid architecture SDR [19].

Another attractive approach to implement SDR is to use multi-core processor architecture. Because of the power wall, memory wall, and ILP (instruction level parallel) wall [5], single core based GPPs may not be able to increase computation speed significantly, and parallel architecture has became the dominate architecture in the industry. Multi-core architecture can use parallelism in achieving SDR's execution speed requirement while still maintaining the same level of flexibility as GPPs. For example, the Cell Broadband Engine (Cell BE) has nine

heterogeneous cores [20], nVidia GPU has 256 cores [21], and Intel has an 80-core CPU [22]. Different signal processing functions can be executed on different cores in a pipeline mode. However, there are still some big challenges for both hardware and software architectures in parallel computing [5, 23], and it is challenging to program in parallel.

On the other hand, high latency tolerant protocols can be used in networks built from SDR nodes with long execution latency [10]. The following is a non-exhaustive list of possible protocols, including changes to existing protocols [10], which could solve the problem: (1) *TDMA*: using a TDMA protocol would solve most of the latency problems, though it requires synchronization among the participating nodes. (2) *Universal Header Coding*: based on a very simple modulation scheme, headers and ACKs are the same through all the different protocols. This way, PHY layer functions can be implemented in hardware and thus a quick ACK or CTS response can be guaranteed. (3) *Delayed ACK and a simple MAC*: SDR execution latency is a big problem in the ACK reply and the RTS/CTS exchange. For the ACKs, we can delay them to a later point in time. We can also adopt a simple MAC like ALOHA protocol [24], instead of CSMA/CA, therefore, avoiding the RTS/CTS exchange.

REFERENCES

[1] E. Blossom, "Exploring GNU Radio," *http://www.gnu.org/software/gnuradio/doc/exploring-gnuradio.html,* November 2004.

[2] http://ossie.wireless.vt.edu/.

[3] "http://sca.jpeojtrs.mil/."

[4] T. J. Kacpura, L. M. Handler, J. C. Briones, and C. S. Hall, "Updates to the NASA Space Telecommunications Radio System (STRS) Architecture," in *Software Defined Radio Technical Conference,* Denver, 2007.

[5] J. L. Hennessy and D. A. Patterson, *Computer Architecture : a Quantitative Approach*, 3rd ed.: San Francisco, CA : Morgan Kaufmann Publishers, 2003.

[6] A. Silberschatz, P. B. Galvin, and G. Gagne, *Operating System Concepts*: Wiley; 7 edition, 2004.

[7] F. Ge, A. Young, T. Brisebois, Q. Chen, and C. W. Bostian, "Software Defined Radio Execution Latency," in *Software Defined Radio Technical Conference,* Washington D.C., October, 2008.

[8] J. Mitola, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Royal Institute of Technology (KTH), 2000.

[9] M. McHenry, E. Livsics, T. Nguyen, and N. Majumdar, "XG dynamic spectrum access field test results," *IEEE Communications Magazine,* vol. 45, pp. 51-57, 2007.

[10] T. Schmid, O. Sekkat, and M. B. Srivastava, "An Experimental Study of Network Performance Impact of Increased Latency in Software Defined Radios," in *WiNTECH: The Second ACM International Workshop on Wireless Network Testbeds, Experimental evaluation and CHaracterization,* Montreal, QC, Canada, 2007.

[11] http://www.redpinesignals.com/.

[12] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice capacity of IEEE 802.11b, 802.11a and 802.11g wireless LANs," *IEEE Global Telecommunications Conference,* vol. 3, pp. 1549-1553, 2004.

[13] "Birth of Broadband," in *International Telecommunications Union*, September, 2003.

[14] J. Neel, S. Srikanteswara, J. H. Reed, and P. M. Athanas, "A comparative study of the suitability of a custom computing machine and a VLIW DSP for use in 3G applications," in *IEEE Workshop on Signal Processing Systems (SIPS)*, 2004, pp. 188-193.

[15] A. Nilsson, E. Tell, and D. Liu, "An accelerator structure for programmablemulti-standard baseband processors," in *WNET2004,* Banff, AB, Canada, 2004.

[16] T. Takano, H. Gambe, and T. Katoh, "Fujitsu's Challenges in Wireless Communications," *Fujitsu Sci. Tech. J.,* vol. 38, pp. 121-133, 2002.
[17] H. Jiao, A. Nilsson, and D. Liu;, "MIPS Cost Estimation for OFDM-VBLAST systems," in *IEEE Wireless Communications and Networking Conference,* Las Vegas, NV, USA, 2006.
[18] P. Athanas, J. Bowen, T. Dunham, C. Patterson, J. Rice, M. Shelburne, J. Suris, M. Bucciero, and J. Graf, "Wires on Demand: Run-time Communication Synthesis for Reconfigurable Computing," in *International Conference on Field Programmable Logic and Applications (FPL),* Amsterdam, Netherlands, 2007.
[19] A. Fayez, F. Ge, A. Young, S. Nair, J. A. Surís, and C. W. Bostian, "Hybrid GPP/FPGA Architecture." vol. US Provisional Patent 61/131,018, 2008.
[20] J. A. Kahle, M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, and D. Shippy, "Introduction to the cell multiprocessor," *IBM Journal of Research and Development,* vol. 49, pp. 589-604, 2005.
[21] http://www.nvidia.com.
[22] http://www.intel.com.
[23] C. O'Hanlon, "A Conversation with John Hennessy and David Patterson," *Queue,* vol. 4, pp. 14-22, 2006.
[24] F. F. Kuo, "The ALOHA System," *ACM SIGCOMM Computer Communication Review,* vol. 25, pp. 41-44, 1995.

7 Limitations of RF Systems and Components in SDR

Although Software Defined Radio replaces much dedicated radio circuitry with more flexible digital processing, it is not possible to replace the RF systems entirely. An increasing demand for flexibility has placed greater demands on the RF components than was previously the case with special purpose radios. Flexibility in operating frequency demands either wide band or frequency agile antennas. Filters must be either eliminated or made frequency agile. Wide band receivers capable of receiving several simultaneous signals demand higher dynamic range to handle the combined energy of numerous interferers.

In this chapter we consider certain aspects of the performance limitations of radio systems between the propagation path and the A/D or D/A converter.

7.1 Key RF Characteristics Considered

Sensitivity is a primary characteristic of radio receivers. Sensitivity is determined by the noise added to a received signal by the receiver circuitry and any resistive losses in the system. Sensitivity may be expressed as noise figure, noise temperature or signal to noise ratio for a specific signal level. Useful sensitivity is limited by external noise in the RF environment.

Antenna Gain quantifies the interface between free space and the radio system. G*ain* is a combination of the directivity and efficiency of the antenna.
System Gain is the summation of the gains of amplifiers and the losses of filters, mixers and other components expressed as a factor or a decibel ratio.

Bandwidth applies to each stage of the SDR as well as being a characteristic of the signals being processed. In SDR the narrowest bandwidth is likely to be the digital signal filter. Bandwidths at RF are most significant in terms of limiting noise and interference. Antenna bandwidths are often small enough to act as filters for the radio.

Dynamic Range has several specific meanings. Most generally it is the ratio between the weakest signal that may be received (noise floor) and the strongest signal that may be received without causing harmful internally-generated interference. The most useful dynamic range definition is the two-tone dynamic range which measures the third order product of two closely spaced signals. Two-tone dynamic range is directly related to third order intercept (IP3).

7.2 Limits of Sensitivity in SDR Receivers

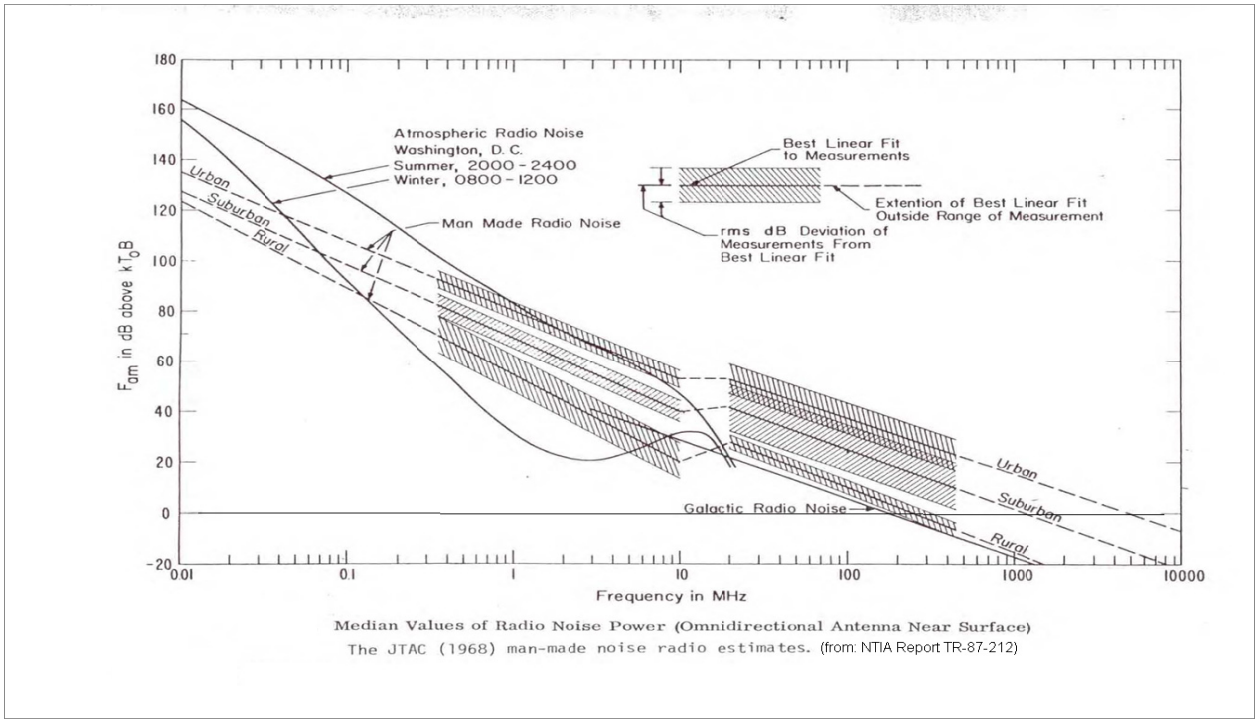


Figure 7.1: Natural and Man-made Noise in the Radio Environment. [1]

Sensitivity is often thought to be the first consideration in radio receivers. Through much of the radio spectrum, however, the ultimate sensitivity of the radio is limited by external natural and manmade noise.

Figure 7.1 shows natural and man-made noise in the radio spectrum from low frequencies to microwave. Noise levels are high in most locations at least up to 1 GHz. In the years since this survey was taken there has been a revolution in computing. Anecdotal evidence indicates that the man-made noise level has increased substantially at frequencies used for computer clocks. Zero dB on the plot is the noise produced by a matched resistive termination normalized to 1 Hz. Factoring in the channel bandwidth gives the expected noise power at a given frequency. For an ideal receiver this value is also the noise floor and, therefore, the limit on sensitivity.

Maximum practical sensitivity is achieved when the receiver noise is substantially less than the sum of all of the external noise sources at a given frequency. Once sufficient sensitivity is achieved further increases in sensitivity results in only small (or no) improvements.

Designing for maximum sensitivity often results in poor dynamic range performance. A better approach is to design for best dynamic range given an acceptable sensitivity. Shajedul Hasan and Ellingson [2] have calculated an optimum sensitivity (expressed as noise figure) for terrestrial

radio circuits at frequencies under 900 MHz. As an example, an optimum noise figure of approximately 15 dB is indicated at 100 MHz for a business environment. For a rural environment a value of approximately 7 dB is indicated. That these values seem high may indicate a tendency to overdesign for sensitivity. Care should be taken to consider the assumptions and application when using this information.

7.3 Limitations of Antennas and Associated Systems

Overview

The goals in antenna development for commercial products focus on size and cost reduction. Laws of physics prevent major breakthroughs in size reduction. Developments are evolutionary rather than revolutionary. The field of antennas is largely mature and old antenna types and solutions work very well in many applications. So, antenna development often centers on adapting known technology and techniques to current applications. This, however, can be challenging at times. There is no magic software package that designs antennas given a set of specifications.

In military applications, performance requirements usually drive designs harder than cost constraints. Smaller size requirements exist in this arena too.

Research and development in antennas currently is heavily directed toward compact and wideband antennas. This is driven by multifunctional, multiband operation where a platform is performing many functions at once, operating in several communication bands, and doing position location and navigation, sensing, receiving broadcast services, etc. Specific applications such as ultrawideband (UWB) and cognitive radio are emerging areas with antenna challenges. The military has had specific needs for years to have an antenna to operate from 2 MHz to 2 GHz. The fact that a good solution has not emerged for this problem is testimony to the difficulty of the task.

7.3.1 Fundamental Limit Theory on Antenna Performance[19]

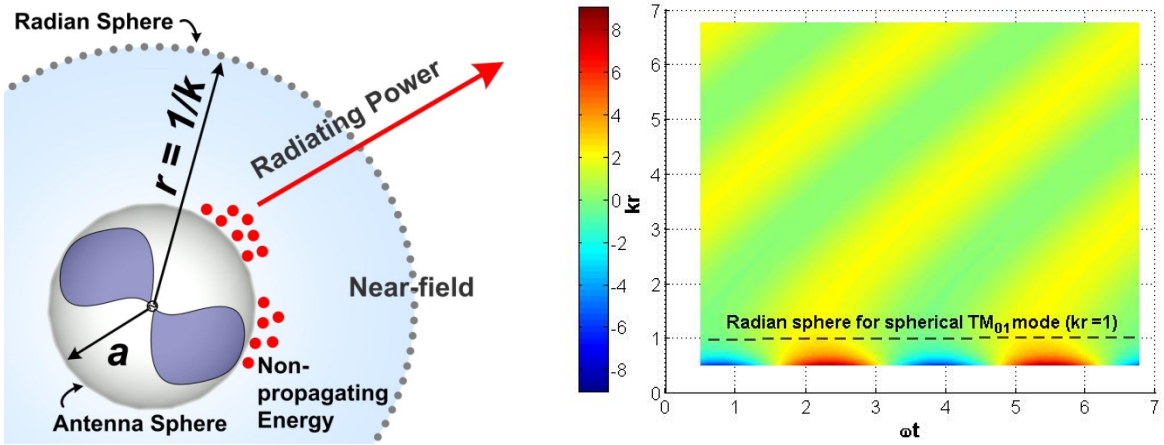
Since Wheeler [3] first introduced the concept of fundamental-limit theory on antennas, there have been many investigations into the theoretical limitations of antenna performance versus size. This theory not only provides a performance limit curve, but also helps to understand the physics of the antenna radiation process, which provides insights for handling interaction issues between the radiating antenna and lossy materials/antennas in proximity distance [4]. As shown in Figure 7.2a, define the *antenna sphere* as the smallest sphere which encloses the antenna structure. Assume that only the fundamental spherical mode (TM_{01}) exists. Then, some of the

[19] This section was contributed by W.A. Davis (wadavis@vt.edu) and Taeyoung Yang (mindlink@IEEE.org) of the Virginia Tech Antenna Group, Virginia Polytechnic Institute and State University, Department of Electrical Engineering.

energy provided by the antenna structure remains in the vicinity of the antenna and does not contributing directly to the radiation process. This non-propagating energy dominates inside the radian sphere, of radius $1/k$, where k is wave number while the radiation process dominates outside the radian sphere. The non-propagating energy is trapped and circulates between the antenna source and the radian sphere. To visualize this process, consider the total transient power radiated by an infinitesimal dipole through a spherical surface at distance r given by

$$P(r,t) = P_{av}\left[1+\cos 2(\omega t - kr - \zeta) + \frac{1}{(kr)^3}\sin 2(\omega t - kr - \zeta)\right] \quad (3)$$

where P_{av} is the average radiated power and $\zeta = \text{atan}(1/kr)$ is the excess phase of the power flow. The normalized total transient power is plotted in Figure 7.2b. The energy storage process is clearly observed inside the radian sphere ($kr < 1$). There appears to be a periodic energy leak on the radian sphere and this leaked energy (radiating energy) contributes directly to the radiated power.



(a) (b)

Figure 7.2: (a) Field regions surrounding an antenna and (b) Total power from an infinitesimal dipole (normalized to average power) [5].

Typically, radiation Q (Q_{rad}) has been used as a qualification factor of the limit theory and is defined by

$$Q_{rad} = \omega\frac{W_{non-radiating}}{P_{rad}} \quad (4)$$

where $W_{non-radiating}$ is the total average non-radiating energy and P_{rad} is the average radiated power. The radiation Q is inversely proportional to 3-dB fractional, instantaneous radiation bandwidth, i.e.

$$Q_{rad} \approx \frac{f_c}{BW_{3dB}} \quad (5)$$

where f_c is a center frequency. Thus, an antenna with a larger radiation bandwidth has a lower radiation *Q*.

There have been efforts to find an exact formulation for minimum radiation *Q*. Chu [6] derived a minimum radiation *Q* expression using circuit models corresponding to spherical-mode wave impedances. Later, Collin [7] derived expressions for the radiation *Q* based on the evanescent energy stored near an antenna, and he also developed the radiation *Q* for cylindrical waves. Fante [8] used the same approach of considering excitation of both spherical TM and TE modes. Foltz [9] and extended the fundamental-limit theory to antennas with a large aspect ratio. Grimes [10] investigated radiation Q with a time-domain approach. Yaghjian and Best [11] provided a convenient formula estimating radiation *Q* from antenna input impedance.

Most recently, Davis *et al* [5] noticed that all previous fundamental limit papers are in error because the previous authors assumed the radiated energy travels with the speed of light all the way from the source. In fact, the energy velocity in the radial direction becomes slower than the speed of light inside the radian sphere. This energy velocity is related to the time delay in converting the stored energy to radiation. Thus, using a different energy velocity in the evaluation of the limit theory provides a different radiation *Q* formula. The resulting *Q* formula of Davis *el al* is lower than other limit theories, which implies that the theoretical limit of achievable radiation bandwidth is lower than what other previous theories predicted.
On the other hand, previous works have been concentrated on electrically, small resonant antennas having minimum *Q*, even though the limit theory was not necessarily limited to electrically, small antennas. Yang *et al* [12] noticed that lower bound of the operational bandwidths of ultra-wideband or frequency-independent antennas are limited by the excited fundamental spherical mode. They proposed limit curves for these antennas in terms of size of antenna sphere and lower bound of radiation bandwidth (3dB cut-off frequency)

The essence of the fundamental-limit theory for antennas is that size, radiation efficiency, and fractional radiation bandwidth are traded off in the design process, and only two of these factors can be optimized simultaneously. Thus, the fundamental-limit theory of antennas provides a theoretical limit to assist in the evaluation of antenna performance in terms of the antenna parameters as well as avoiding the search for an antenna with unrealistic performance parameter values. Unfortunately, efforts over the past half century have only focused on the accuracy of the theory. As shown in Figure 7.3, most of the conventional antennas are not even close to the limit curves.

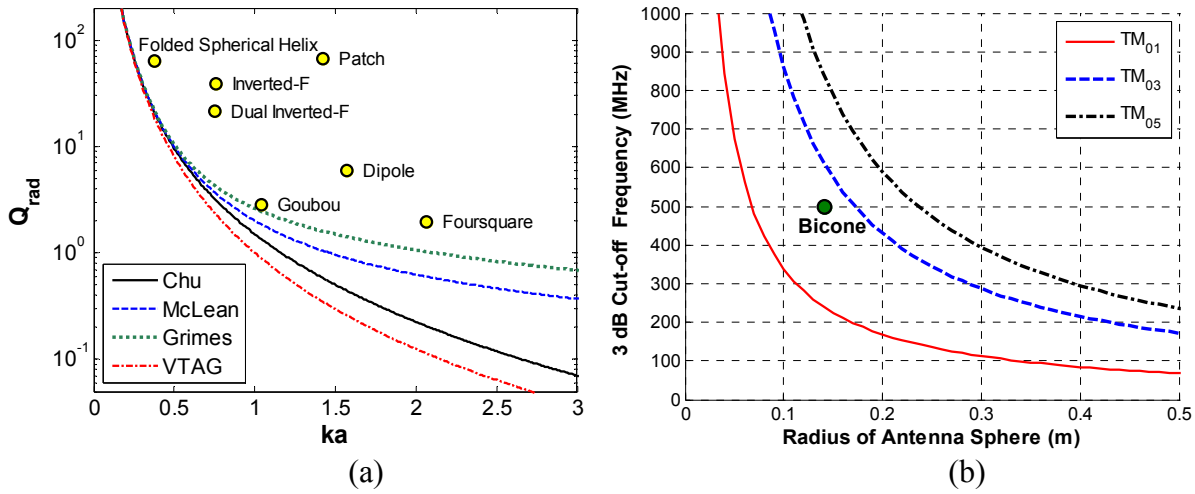


Figure 7.3: Limit curves of various fundamental-limit theories and comparison with conventional antennas – (a) For resonant antennas [5] and (b) For ultra-wideband or frequency independent antennas [4].

Rather than arguing with the accuracy of the limit curves, more productive research discussions on the fundamental-limit theory would be based on the following questions:

1. Why are conventional antennas not close to the limit curve?

2. How can we make the conventional antennas close to the limit curve?

In addition, as implied by Yang *et al* [4], an antenna close to the theoretical limit has less interaction with lossy materials or other antennas in proximity distance. Therefore, the antenna close to the limit will have improved radiation efficiency and less interaction with other adjacent antennas. This aspect will be useful in co-site planning.

Note that any material loading techniques (for example, metamaterials or dielectric-magneto materials) cannot change the theoretical limit. Sometimes, these materials help to bring conventional antennas close to the limit curve. But these approaches create other problems, like increased loss, reduced bandwidth, and increased weight. If these issues are resolved, the material loading techniques will provide the most convenient way to bring the existing conventional antennas to the theoretical limit without changing the original antenna structure significantly.

7.3.2 Practical Antenna Limitations

Antenna size reduction by reactive "loading" has been practiced since the earliest days of radio. All known forms of loading have been used. Examples include inductive, capacitive, "linear" (transmission line), dielectric and magnetic (materials).

Despite the many claims to the contrary, all known forms of loading are subject to the restrictions of the fundamental limit theory of the previous section. The three-way trade-off between efficiency (gain), bandwidth (Q) and size (in terms of wavelength) remains the most important limitation in passive antennas. Auxiliary considerations may also become important when making an attempt to reduce antenna size. Extreme cases may be encountered in electrically small antennas with low efficiency, narrow bandwidth and high voltages and/or currents.

Low efficiencies are common in electrically small antennas. Losses of several dB with respect to a full size antenna are common. Cases having 20 dB or more loss are not unknown. Inefficiency may either be due to an attempt to operate the antenna at sizes smaller than allowed for 100 percent efficiency or from causes unrelated to the fundamental limits.

As an antenna becomes smaller and the Q rises there is an accompanying increase in the currents and voltages in the antenna structure. Heating due to resistive losses in the antenna can result in problems like tuning shift and even failure of supporting structures. High voltages may be encountered as the small antenna often resembles a Tesla Coil. Dielectric breakdown in the antenna structure can result in permanent failure of the antenna. Corona discharge and arcing limits efficiency and may occur even at low power. Painful RF burns are a hazard to the operator and can occur at powers of only a few watts. This suggests that there are practical limits related to materials. It would be difficult to cover this in general terms due to the many different antenna structures, loading methods and materials encountered.

Small antennas having low losses invariably have a high Q. High Q implies a narrow bandwidth. Extremely narrow bandwidths may be encountered. It is possible to design and build antennas that have insufficient bandwidth to pass the desired modulation. Occasionally this is true even for narrow band modulation modes. Modulation bandwidth represents an important limit because it directly affects the usefulness of high Q – small antennas.

7.3.3 Antenna Tuners

An “antenna tuner” (or transmatch) is device which provides a conjugate match between a radio and an antenna [13]. Combinations of inductors and capacitors are adjusted to both cancel reactance and transform the resistive component of an antenna’s input impedance to a value preferred by the radio designer (usually 50 Ohms). It is interesting that only three adjustable circuit elements (e.g. two capacitors and one inductor) can handle both functions over widely varying antenna conditions. Implementing a tuner with practical components can be a challenge, especially when covering a wide frequency range. A tuner is a narrow band device which should be automatically adjusted by the SDR.

Automatic tuners are readily available in the HF region of the spectrum. HF tuners usually use high-Q inductors and capacitors. The tuner is often used to force feed a narrow band antenna like a monopole over a wide frequency range. Antenna feedpoint resistances (real part) may

vary from a fraction of an Ohm to thousands of Ohms. Feedpoint reactances may also have extreme values and be either sign. Actual impedance values are strongly dependent on frequency. At extreme antenna impedances the Q of the tuner becomes quite high resulting in narrow bandwidths, high voltages and high currents. High circuit losses will occur unless extremely good components are used.

At VHF practical tuners make a transition from lumped circuit components to transmission lines. Transmission line tuners are used almost exclusively at UHF and above. A "three stub tuner" is a classic example of a general purpose tuning device. Transmission line tuners often require the use of sliding contacts in the adjustable components. Sliding contacts are often a source of trouble in tuners due to the high circulating currents that often exist in the circuit. Percentage frequency ranges and permitted impedance ranges of VHF and above tuners are less than for HF tuners. Fortunately, VHF and above antennas are often are relatively large which results in more favorable range of feed point impedances.

Given the difficulty of producing a tuner which covers all of the possible impedances at the terminals of a simple antenna it is often advisable to integrate some tuning elements directly into the antenna itself. If a discrete tuner is included it should be located as close to the antenna as possible in order to reduce losses.

If a wide instantaneous bandwidth is desired for applications such as UWB or simultaneous operation on several dispersed frequencies, we should choose a naturally wideband antenna to do the job.

7.3.4 Non-Foster Matching

"Non-Foster matching" has been proposed by numerous authors as a solution to the antenna matching problem. For any given antenna, the terminal impedance varies with frequency in a manner that is not compensated for by any combination of realizable passive components. This is a consequence of Foster's Reactance Theorem [14]. It is possible, however, to construct a compensator using a gyrator or a "negative impedance converter" which synthesizes the negative inductors or negative capacitors needed to work around Foster's Reactance Theorem [15]. Simulation has shown that the principle may be extended to create multi-octave wide band arrays [16].

Circuits used to implement the negative impedance converter use transistors or other active devices in feedback circuits to generate negative impedances. Attempts at practical implementation of non-Foster matching have often been frustrated by instability (oscillation), noise (in receivers) and power limitations in transmitters. Some positive results have been reported, however [17].

Non-Foster matching techniques may become an important tool in the radio engineer's tool box. Caution should be exercised, however, to consider the limitations of the technique. Since active

devices are used to synthesize non-realizable passive components, noise will be contributed when compensating reactances for receive antennas. In transmit applications; the negative impedance converter takes the form of a power amplifier. The ability of the circuit to compensate for reactances in power generating applications is limited by the voltage and current capabilities of the active device.

7.4 Limitations of SDR Receiver Architectures

There are several different receiver system architectures in use today. The receiver architecture chosen profoundly affects the amount of digital processing required. Dynamic range and power consumption are also impacted by the choice of receiver architecture. The number of independent frequencies that must be received simultaneously also strongly impacts the choice of receiver architecture.

A receiver which can process several independent signals at once is called *multiple carrier receiver* [18]. A *single carrier* receiver, however, is only capable of receiving one signal at a time (unless provided with additional hardware for each signal).

In this section we consider the limitations of the most common choices of receiver architectures in use today.

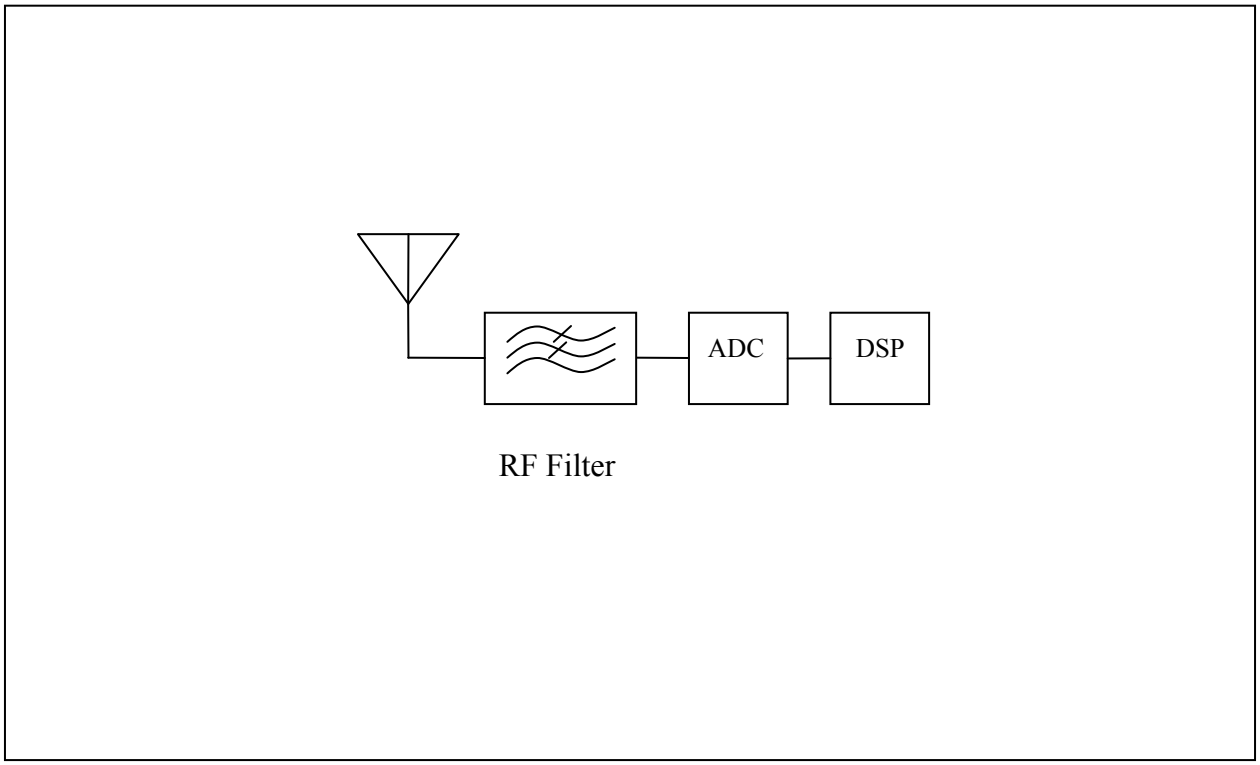


Figure 7.4: "Ideal" Receiver.

7.4.1 "Ideal" Receiver

The "ideal" DSP based receiver in Figure 7.4 has only the antenna and low pass (or bandpass) anti-aliasing filter ahead of the analog to digital converter. This type of receiver is often called a *direct sampling receiver*. This type of receiver has a great deal of flexibility. There are practical performance limitations, however. As our previous chapter indicated, the analog to digital converter must have a large number of bits in order to achieve an acceptable noise figure. At present this scheme can be used at lower frequencies. At VHF through microwave it is impossible to satisfy sensitivity, dynamic range and Nyquist requirements at the same time. An example describing a 12 bit ADC with a 33 dB noise figure appears in [19]. Obviously, a RF amplifier is required!

Additional difficulties present themselves, however. Since at least the input stages of analog to digital converters contain analog circuits, they have limitations in linearity. Given the high dynamic ranges required from radio receivers (80 dB+) and the large number of signals in the received radio spectrum, even the tiny nonlinearities in a high quality receiver swill cause intermodulation components to be generated. The intermodulation components can fall on the desired receive channel creating interference that, once produced, cannot be distinguished from an actual received signal.

Our example "ideal" receiver is particularly vulnerable to intermodulation production since a large number of signals almost certainly are present at the input of the ADC at any given time. In addition to the third order intermodulation components that plague narrow band receivers, wide band receivers must contend with second order intermodulation products. With wide band receivers, even a single strong undesired signal can produce interference at harmonics of its frequency. For that reason alone, a conventional RF bandpass filter with a passband of less than an octave is required for good performance in the wideband radio receiver.

Given the practical requirements, the receiver now looks a little more complicated. Figure 7.5 shows a direct sampling receiver where we have added band pass filters and a low noise amplifier. As mentioned above, the bandpass filters should be designed to suppress harmonics and second order products. A number of filters may be required to cover large frequency ranges.

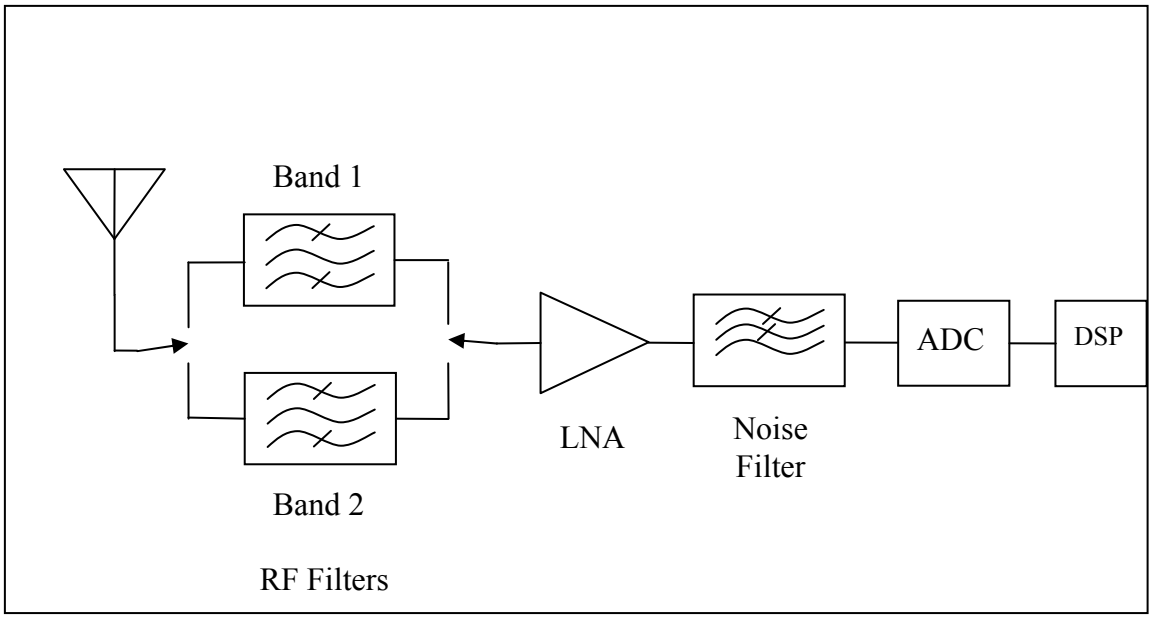


Figure 7.5: Practical Direct Sampling Receiver

Even with the bandpass filters the low noise amplifier and ADC are called on to process many signals simultaneously. Often the desired signal is orders of magnitude weaker than other signals in the passband. A receiver design must manage the operating range of the amplifiers and ADC so that all incoming signals are maintained in the linear range. Amplifier gain must be sufficient that the weakest desired signals (near the noise level) are not suppressed. That is usually taken to mean that "several" of the least significant bits of the ADC will be occupied by the incoming noise. At the same time the strongest signals must not produce significant intermodulation products or saturate the ADC.

The most significant difficulties with this type of receiver are dynamic range, ADC speed and available DSP processing power. The advantages are flexibility and the ability to process more than one signal at a time.

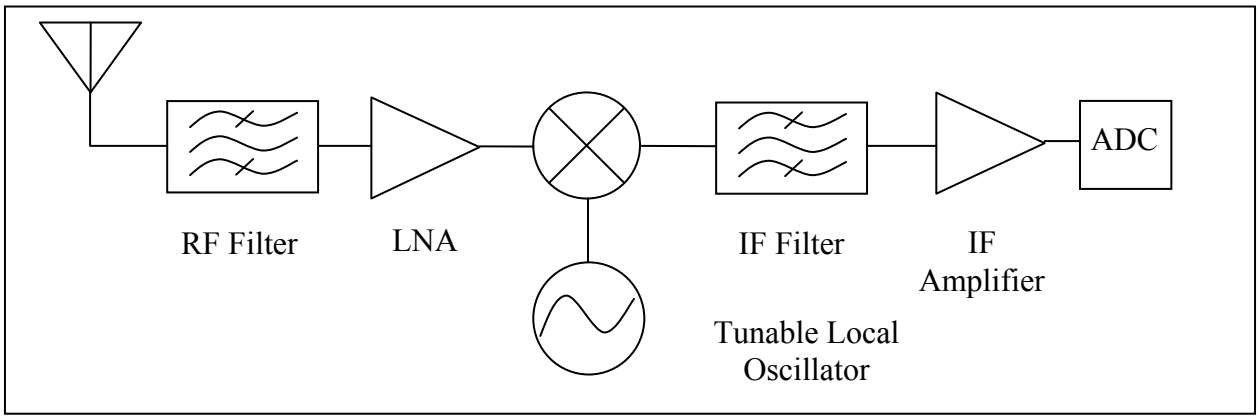


Figure 7.1: DSP Based Superheterodyne Receiver

7.4.2 Superheterodyne Receiver

The most common conventional receiver architecture is the superheterodyne. Superheterodyne receivers make use of a mixing process to convert the desired signal to another, intermediate, frequency for amplification and filtering before converting to baseband. DSP may be applied directly at baseband. A more effective approach, however, is to sample the signal at IF and make the final conversion to baseband with DSP.

In a DSP based superheterodyne, shown in Figure 7.6, signals received at the antenna are first filtered to limit signals to the desired frequency band. Amplification follows to increase the signal level and set the noise figure. The mixer converts the received frequency either up or down to an intermediate frequency (IF) by multiplying the incoming frequency band with the local oscillator frequency. Following the mixer the IF filter selects the intermediate frequency of interest. Then the IF signal is amplified to levels appropriate for conversion to digital. The digitized IF signal is converted to baseband in the digital processor.

This type of receiver is capable of achieving high dynamic range due to the selectivity provided by the IF filter. Both sampling rate and DSP processing are reduced by the narrow bandwidth of the IF filter. Good frequency agility may be had when the local oscillator is a frequency synthesizer under software control. The tremendous advantage of the superheterodyne receiver architecture is the abundant “free processing” provided by the IF filter.

7.4.3 Block Down-Conversion

If a fixed frequency local oscillator is used along with a wide band IF filter in the Superheterodyne receiver a “block” of frequencies may be converted to IF and passed to the ADC. The DSP can filter to select the desired signal from the IF and then down convert it to baseband. As an example, an IF bandwidth 20 MHz centered at 70 MHz could be chosen. Since no LO tuning is required in this type of system the DSP operates as it would in the direct sampling receiver case.

The principle of block down conversion may also be used along with the superheterodyne receiver or the direct conversion receiver (described below). Such combinations are referred to as multiple conversion receivers.

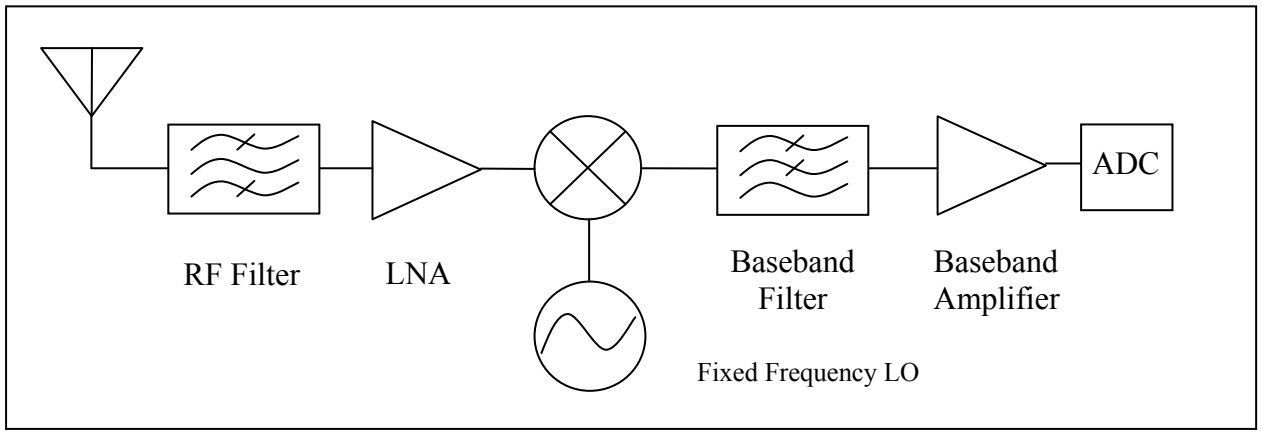


Figure 7.2: Direct Conversion Receiver

7.4.4 Direct Conversion Receiver

The direct conversion receiver shown in Figure 7.7 is similar to the single conversion receiver except that the local oscillator frequency is chosen to move the received signal directly to baseband. The last filter function is now lowpass in order to allow for signal components near zero frequency. A complex (I-Q) conversion is usually performed so that the image can be eliminated or reduced.

A significant advantage in the direct conversion receiver is that baseband filtering is substituted for the IF filtering. Baseband filtering can be accomplished by active filtering with operational amplifiers or simple passive networks which may be easily adjusted to accommodate different bandwidths. Usually a complex conversion is performed since it is usually not otherwise possible to filter out the image frequency. Image rejection is limited, however, by the balance of the converter components. Typically 35 dB image rejection (or worse) is obtained.

With the direct conversion receiver gain distribution becomes critical. Low frequency "1/f" noise in the baseband amplifiers cause a problem with sensitivity since the response extends down to DC. Increasing the RF amplifier gain can increase the sensitivity but reduces the dynamic range. Since most of the gain in this receiver occurs at baseband we may also have feedback problems unless great care is taken with isolation between input and output. Power supply voltages must be filtered between the gain stages to eliminate feedback. Power supply noise filtering also becomes critical to eliminate hum and other noise superimposed on the supply voltage.

While the direct conversion receiver provides the designer with numerous difficult challenges it may be the lowest cost choice for a receiver covering a wide frequency range. Analog active filters may be used to filter the baseband signals before sampling. Sampling of the filtered baseband signals requires a relatively low sampling rate. DSP processing requirements are also

low. Because of this the overall receiver power requirements and cost are low. The direct conversion approach is commonly found in single chip radios for handheld and mobile use where cost and power consumption are primary considerations.

7.4.5 Comparing Receiver Architectures

While comparing the receivers described above we see two general categories emerge. Those categories are differentiated by their ability to receive either one or more than one carrier (signal) at a time. Receivers such as direct conversion and superheterodyne belong in the *single carrier* category. The single carrier category of receivers is appropriate where only a single channel is in use at any given instant. In contrast, the direct sampling and block down conversion receivers perform the final conversion to baseband digitally. They fit into a truly wideband, *multiple carrier* category where it is possible (in principle) to receive multiple signals in one hardware system.

With single carrier receivers the actual receiver bandwidth may be either wide or narrow, depending on the filter settings. Any type of modulation can be accommodated if complex sampling is used. Final filtering and processing of the baseband signal is done in the digital domain. With this category of receivers the sampling rate and demands on the DSP system are low because of the filtering provided by an IF or baseband filter. Analog multipliers used in the direct conversion and baseband sampling superheterodyne receivers are subject to noise problems as well as drift, imbalance and DC offset problems. Software multipliers used in IF sampling may be made as good as the numerical precision of the digital signal processing allows. This is advantageous but increases demands on both the ADC and DSP since higher frequency sampling is usually necessary to accommodate the incoming IF frequency. DSP has more to do in this case due to the digital conversion and additional filtering functions. Because of the deficiencies of analog multipliers, receivers having IF sampling are superior to those with baseband sampling.

If multiple independent signals are to be received at a given instant the single carrier design is insufficient as each signal needs to be separately converted to baseband. Multiple carrier receivers provide for this when the incoming signal is sampled at either RF or IF frequencies. Simultaneous conversion of several signals to baseband may be accomplished numerically by the complex multiplication of the incoming spectrum by a numerically generated local oscillator at or near each desired signal frequency. This type of receiver is important in base station systems where many incoming transmissions may be received at the same time on adjacent frequencies. This type of receiver also has the advantage of removing the critical conversion to baseband from the analog domain to the digital domain where a more ideal function may be implemented numerically.

Disadvantages of multiple carrier receivers include difficulty meeting dynamic range specifications and high processing requirements. In addition, obtaining sufficient dynamic range in wide band, multi-carrier receivers is challenging. Such receivers have wide band amplifiers

from RF through the end of the IF chain. Since intermodulation products increase in level faster than the desired signals, systems where the potential intermodulation precursor signals are amplified to high levels will have a significant disadvantage. Systems having significant filtering prior to amplification will eliminate much of the intermodulation simply because the precursor signals are absent. Suppression by filtering of the precursor signals at any point in the receiver chain essentially eliminates the generation of intermodulation products in the following stages.

7.4.6 Tuning Range in Mixer Based Systems

Mixers are a primary component of radio systems. Their function is an approximation of a time domain multiplication of a signal with a local oscillator (LO). While by definition this is a non-linear function, the desired output component is a replica of the input signal converted to another frequency. So the mixer is treated in the system as if it were a linear component. In addition to the desired response of the mixer there are many undesired responses and outputs. Generally it is necessary to suppress the undesired responses and outputs with filters.

Undesired mixer outputs include (but are not limited to) leakage from the local oscillator, harmonics of the local oscillator, harmonics of the input signal, leakage from input signals and mixing products of undesired signals with the local oscillator.

Usable RF bandwidth is limited by spurious responses (spurs) of the mixer stage. Spurs are a natural product of the mixing process. Responses of a mixer may be found by: $f_{if} = Mf_{rf} \pm Nf_{lo}$. The primary mixer responses occur at M=1 and N= any value. The desired response is usually M=1 and N=1. Exceptions are found in test equipment such as antenna range receivers and spectrum analyzers which use N≥1 mixing. Occasionally, "low cost" receivers will also be produced to use N>1 mixing. N>1 mixing is characterized by increased mixer loss. Here we only consider N=1 as a desired mixing component. Other responses are undesired. Balanced mixers are usually designed so that even order N spurs are suppressed. This can reduce the signal amplitude of a spurious response by 20 to 30 dB with respect to the N=1 component. This is a large help in mixer spur reduction. Further reduction in the spurious response must be provided by the filters preceding the mixer.

Relative amplitudes of the responses are given in by Rhode and Bucher for a typical double-balanced passive mixer [20]. The desired response is usually for M=1 and N=1. Either the sum or difference may be selected as the desired response. The other response is usually suppressed and is called the "image". For values of M=1 and N=2, 3 and 4 typical relative response values are 35, 15 and 37 dB less than the M=1, N=1 case. Higher order values are less troublesome with all M>1 cases having losses of about 60 db (or greater) with respect to the M=1, N=1 response.

Required suppression of the image and spurs depends on the radio type and operating goals. Radio designers must carefully consider where the undesired responses occur in the radio frequency spectrum. For example: it would be highly undesirable for a land mobile radio

receiver spur or image to fall on the frequency of a local television transmitter. Even with a high degree of suppression the television transmission could cause interference to the desired signal. Various authors have suggested different values for required image rejection. Rhode and Bucher suggest a rejection specification of greater than 80 dB for both image and feed through of IF signals from the antenna [21]. Kraus, Bostian and Raab suggest a 50 dB minimum specification for image rejection [22]. Neither value is universally right or wrong. A base station radio receiver may require a higher image rejection while less may be appropriate for a hand-held radio.

Mixer spurs and images represent a limitation (of a sort) for SDR. In receiver types based on either the superheterodyne or block down conversion approach it is necessary to limit the bandwidth of the RF filters to ensure that the image and higher order spurs are suppressed. Tuning bandwidth is limited so somewhat less than an octave for a fixed filter approach. Tuning wider frequency ranges requires either switched filters or a continuously tuned filter. For the block conversion receiver where multiple carrier reception is required this can be an important limitation.

7.5 Limitations of Automatic Gain Control (AGC) in SDR Receivers

An AGC circuit is incorporated in almost every radio receiver. The primary function of AGC is to reduce the amplifier gain to keep the level of the desired signal either constant, or within the linear operating range of the receiver system. Typically, at zero signal level the system gain will be at maximum.

Out of band signals present a conceptual difficulty for direct sampling and block down-conversion SDR receivers where a wide bandwidth is maintained before the ADC and DSP. If the AGC signal is derived only from the desired receive signal and is used to control the gain of wide band amplifiers ahead of the ADC then strong out of band signals may overload the amplifiers. It is important, therefore, to limit the gain of RF and IF components so that overload doesn't occur due to undesired signals. In order to do this it may be necessary to either arbitrarily limit the gain of the amplifier stages or derive the AGC feedback signal from a point prior to any digital filtering.

Latency in AGC feedback derived from the DSP processor can cause instability in the AGC loop. Unless the time delay caused due to DSP processing is quite short the AGC will not be able to follow short period fading due to propagation of the RF signal. AGC loop gain will have to be low to avoid oscillation due to the phase shift resulting from the time delay. This is a traditional control system problem.

7.6 Receiver Dynamic Range Limitations

Dynamic range is an issue in SDR because the radio environment contains many undesired signals in addition to the signal that we desire to receive. Some of those signals may be

hundreds or thousands of times stronger than the desired signal. Given the high gain required to bring our desired signal to a detectable level, other strong signals may drive the amplifier, mixer or ADC into their non-linear region.

In all systems third order products may cause interference since they may be produced by two adjacent channel interferers. Second order intermodulation products only occur when one or both interferers are located far in frequency from our desired channel. Second order products are usually only a problem in wide band amplifiers. Higher order products exist but are usually less of a problem since the second and third order products appear first.

A SDR receiver must be capable of amplifying signals near the ambient received noise level to levels sufficient to drive the ADC. The output of the amplifier must be able to produce enough voltage (or power) to drive the ADC to full range. Any distortion of the incoming waveforms causes discrete intermodulation products to form. These products are virtually indistinguishable from real signals in the receiver output. Any intermodulation products that fall on our desired signal frequency will cause interference.

When we depend entirely on DSP filters to separate the desired signal from many other signals spread over a large portion of the frequency spectrum, as in the direct sampling receiver, the amplifier and ADC must have extremely good dynamic range to avoid the generation of many undesired intermodulation products.

Poberezhskiy [23] develops an explanation for this by first assuming that the number of signals in the receiver signal path increases linearly with frequency. Each pair of signals can produce second order products ($f1\text{-}f2$ and $f1+f2$) and third order products ($2f1\text{-}f2$ and $2f2\text{-}f1$). As the number of incoming frequencies increase the number of intermodulation combinations increase much more rapidly (e.g. $2f1\text{-}f3$, $2f1\text{-}f4$, $2f2\text{-}f3$). Therefore, as the bandwidth of the receiver increases, the chances that intermodulation products will fall on our desired signal increase rapidly.

Superheterodyne receivers share this problem with the direct sampling receiver. The problem occurs in the RF and early IF stages where the stage bandwidth is wide. Generation of intermodulation components can be less in the wide band portions of the superheterodyne because the gain is kept low. Final amplification then occurs after the narrow band channel filter. The channel filter tends to shield subsequent stages from intermodulation generation by stripping off the interferers before they are able to generate the intermodulation products.

Poberezhskiy [23] develops an approach for estimating the required dynamic range for a SDR receiver. His method depends on the development of a statistical model of the interference where the number of interferers is proportional to the bandwidth seen by the amplifiers and other stages. The required dynamic range is found where the number of significant intermodulation products is small in terms of the number of interferers. Poberezhskiy also shows how the limit of dynamic range is closely tied to the power consumption of the amplifiers, mixers and sampler

stages.The tuner study [24] develops an empirical relation between the bias power required and the dynamic range. Baltus also formulates the relation between power for amplifiers [25] and multistage [26] systems. These studies taken together emphasize the desirability of limiting (by passive filtering) the bandwidth processed by the receiver RF stages.

7.7 Limitations of SDR Transmitters

Transmitter RF systems take on the same forms as receivers – except in reverse. There are a few notable differences, however. The signal level distribution is quite different since there is increasing power in the transmitter as we move forward toward the antenna. Power levels may easily be computed at each stage since the starting level is known. By computing the gains and power levels at each amplifier stage an optimum power budget can be arrived at. Most amplifier stages of a transmitter are operated near their maximum linear power levels.

In receiver systems spurious signals are only a matter of performance specifications. In the transmitter, however, it also becomes a regulatory and interference problem. Usually a transmitter will have to be tested by a certifying laboratory and registered with the proper government agency. Care must be taken in the design to ensure that the transmitter maintains specification throughout its entire range of frequency, power level and modulation modes.

7.7.1 Transmitter Noise

Wideband "white noise" is produced by transmitters at frequencies both near and well away from the intended transmitting frequency. Control of transmitter noise is necessary to reduce interference to nearby receivers operating at other frequencies. Noise may be subdivided into the phase noise or amplitude noise categories.

Phase noise is largely a function of the system oscillators and is primarily a problem at frequencies quite close to the intended transmitted frequency.

Amplitude noise is present at both far out and close in frequencies. Transmitters often employ bandpass filtering which attenuates amplitude noise outside of the passband but provides little attenuation in band. Even with careful low noise design a transmitter is likely to interfere with co-site receivers. The classical solution is to provide either a narrow band filter tuned to the transmitter output or a notch (reject) filter tuned to the receiver frequency. Unless the filter is frequency agile, however, this is not an entirely satisfactory solution for SDR. Furthermore, narrow band filters are likely to have several dB of insertion loss, wasting transmitter power.

Further development needs to be done in the realm of transmitter noise control. Frequency agile bandpass and notch (or band reject) filters are needed which also have low loss and high tuning rates.

7.7.2 Transmitter Efficiency

While high efficiency power amplifiers, such as class C, D and E are available for constant envelope types of waveforms, lower efficiencies are a problem for waveforms requiring linear amplification.

High efficiency amplifiers use resonant circuits to reconstruct the original waveform. The resonant circuits limit instantaneous bandwidth to some extent. For frequency agile systems the more important limitation may be the need to change filters to accommodate various frequency bands.

7.8 Limitations of RF MEMS Switches

MEMS (Micro Electro-Mechanical Systems) RF switches are finally coming onto the market. At least two companies have published data sheets and appear to be shipping components commercially[20] [27] [28]. Other companies probably are sourcing the devices at this time for military purposes.

There are several different types of MEMS switches. Metal to metal contact switches operate in much the same manner as conventional switches. They may be designed for series or shunt operation. Capacitive switches do not provide a direct metal to metal contact but rely on the motion of a relatively large metallic surface to change significantly the capacitance between two surfaces. Capacitive switches are frequency dependent as the high capacitance state must present a low reactance to the circuit while the low capacitance state must look like a high reactance. For a given switch this results in a broad but limited frequency range.

7.8.1 Contact Type MEMS RF Switches

Contact type MEMS switches have a large "wow factor" due to amazing specifications. A review of the specification sheets reveals high isolation, low loss, fast switching times and excellent VSWR. Device bandwidth is from DC to the GHz region. Un-switched (static) power ratings of several watts are available. Package sizes are comparable with solid state devices but will probably become better with time. The key deficiency, however, is the "hot switching" specification.

"Hot switching" is the change of a switch position while signal power is applied. Hot switching has always been an issue with mechanical relays and switches of any sort. Arcing, pitting, oxidation and even welding of contacts lead to switch failure. Repeated switching under power eventually causes the contacts to degrade to the point of failure. A gross overload of a switch may cause instant failure. Hot switching of MEMS switches seems to be a problem at even low power levels.

[20] It is reported that TeraVicta ceased operations early in 2008.

The power levels at which a switch may be used under hot switching conditions are typically much less than the un-switched levels. For MEMS RF switches this is true in the extreme. TeraVicta Technologies specifies a hot switching level of 1 milliwatt for an example device while Radiant MEMS specifies hot switching to begin at 0.1 milliwatt for one of their devices [29]! The example TeraVicta Technologies data sheet has a de-rating curve showing decreasing switch life for hot switched cases to a level of 10 milliwatts. These hot switching values are probably typical for this type of device.

Most authors seem to avoid discussion of the hot switching problem by assuming that switching is always done at zero power level. Rebeiz discusses hot switching, however, and attributes contact degradation to pitting of the contacts during the switching event [30]. Hot switching should be considered with respect to the potential power levels that could be received at antenna terminals in adverse conditions.

Although the usual received signal levels at the antenna terminals of a radio receiver are quite small, (usually less than -30 dBm) undesired signals from nearby transmitters may be much larger. One example would be on base stations where multiple receivers and transmitters are located on the same tower. Another example of high signal levels would be in the case of two mobile units parked in adjacent spots. The case of mobile units in adjacent parking spots is estimated in the appendix for a transmitter power 100 Watts at a frequency of 150 MHz and an antenna gain of 3 dBi. Free space far field is assumed. A received power level of 0.6W is found for this case. This does not compare favorably with the 10 milliwatt hot switch rating of the MEMS switch. The MEMS switch under consideration may fail if switched at this power level. In the field, failure of contact type MEMS RF switches would probably manifest themselves as random events when used in mobile service.

The DC withstanding voltage rating for open circuit MEMS switches is an interesting matter given the small dimensions of the switch. Typical switch contact separations are on the order a few micrometers. At this distance there is a departure from the breakdown voltage predicted by Paschen's Law. Between smooth conductors there is a constant breakdown voltage of about 340V at spacings of about 1 to 5 micrometers [31]. This compares favorably with the voltage rating given by manufacturer's specification sheets [29].

It seems possible from the dimensions of the contacts that the basic MEMS RF switch may be operating at or near its theoretical ratings. Further improvements may require a different approach like combining multiple contacts or scaling up the devices to somewhat larger dimensions. At this point, however, the contact type MEMS RF switch is not recommended for use except in a controlled environment.

7.8.2 Capacitive RF MEMS Switches

MEMS switches which act as two-position variable capacitors have been developed. The capacitive type switches do not suffer from the hot switching problems of the contact type switches Capacitive switches have been demonstrated that operate to power levels of 6 Watts in the 8-18 GHz frequency region [32]. A limitation does occur due to "self actuation" under high operating power (or voltage) [33].

Another limitation of the capacitive type switches occurs due to the limited difference in capacitance between the on and the off state. A high capacitance is desirable in the on state so that the reactance due to the series switch is small. Conversely, a small capacitance is required for the off state. The small capacitance must be such that the series reactance will be quite large. The ratio between the on and off state capacitance defines the quality of the switch. The actual values of the capacitance in each state set a limit to the useful frequency range of the device. The maximum capacitance is limited by the "real-estate" available on the die.

Although the capacitance values on a practical MEMS device are limited they may be well suited to some applications. TR switching and filter switching in the GHz frequency range may be practical applications for these devices.

7.9 Implications of Limits and Proposed Solutions

7.9.1 Sensitivity in SDR Receivers

Sensitivity is a primary limitation since noise is the ultimate hard limit. While galactic and thermal noise sources are well known and stable, man-made noise sources are highly variable with time and location. In much of the spectrum the practical limits of sensitivity exceed the requirements set by ambient noise levels. In the frequency range below about 1 GHz careful engineering design must be weighted toward dynamic range rather than sensitivity.
The known methods of dealing with the sensitivity limit include:

- Use narrow bandwidth or other filtering
- Use higher transmitter power
- Use higher directivity in transmit and/or receive antennas.

7.9.2 Antennas and Associated Systems

Antennas take on many forms depending on frequency and function. In section 7.2.2 we demonstrated the three way tradeoff between size, efficiency and bandwidth. Despite a large amount of effort to produce efficient electrically small antennas, there are no known antennas that exceed the limits. It is possible, however, to make inefficient antennas having smaller size

or wider bandwidths. Marginal improvement may be had by exploiting orthogonal radiation modes.

Conventional active antennas are quite effective when used for receiving. While they appear to violate the limits they actually side step the issue. Active antennas deserve continuing attention as a means to avoid the size limits.

While there are many clever antenna techniques to be exploited, there is often no substitute for allocating a sufficient volume of space for the antenna.

7.9.3 Receiver Architectures

While the direct sampling SDR has a great deal of flexibility it may prove too difficult to use for highest performance. Where high dynamic range is required, especially at VHF and above, engineers have fallen back on the superheterodyne architecture. [23] [24] High sampling rates and demands for processing also make the direct sampling architecture expensive to implement. Where power consumption is a consideration superheterodyne and direct conversion architectures have the advantage.

Tuning range for receivers is also an issue since it is not feasible to eliminate filters from the radio system due to residual nonlinearities in amplifier and mixer stages. Tuning range for fixed filter systems is best made small. A large number of filters would be required to cover the frequency range between 0.1 and 1.0 GHz with narrow band filters. Practical (miniature) filter banks or better tunable filters are needed for radios [24].

Except for their physical size and power requirements YIG filters [34] would be good candidates for tunable receiver filters. The simple current drive interface of the YIG filter would be convenient for SDR processor control.
SAW filter banks [35] are potentially a useful technology for dividing the frequency spectrum into sections for processing in conventional receiver circuits. SAW filter banks may take the form of a network of individual SAW filters or an integrated device with a single input and multiple outputs.

Filter banks may be used either in the RF or IF sections of a receiver. If used in the RF section of a receiver, the SAW filter bank serves as a diplexer or circuit for dividing the incoming signals between several receivers.

SAW filter banks could also be useful following a mixer stage to divide the IF into several broad bands. Each IF band would then be sampled and processed individually using low speed A/D converters and processors.

7.10 SDR WISH LIST:

- A flexible RF switching arrangement that accommodates several antennas and several receiver inputs along with several transmit outputs. The switching network should allow for half-duplex or full-duplex communication with different or the same antennas. It should also allow for selecting different antennas having differing frequency ranges.
- A flexible RF filtering arrangement incorporating electronically adjustable (or switched) center frequency along with adjustable (or switched) bandwidth. It is desirable that the filter system is coordinated with the switching network in order to allow for a duplexing function when both transmit and receive operate at the same time. It should also allow for several channels of transmit and receive to be used simultaneously.
- Automatic antenna tuning for transmit and receive.
- Transmitters and receivers having several channels each to provide for diversity, beam forming, MIMO and independent operations under software control.
- Antennas having several inputs, each corresponding to a distinct directional beam. This antenna would operate similar to an array of elements fed by a beam forming network (e.g. Butler network). In this case, however, the beam forming is inherent in the antenna structure itself.
- Small size antennas that meet the antenna fundamental limit curve.
- Lower power, higher speed, higher resolution Analog to Digital Converters (of course).
- Electronically adjustable IF filters.
- Better switches!

Here are items we expect to improve with time:

- RF Low Noise Amplifiers (good and getting better).
- Passive Mixers: a mature technology may make small incremental improvements.
- Active Mixers: a somewhat mature technology that keeps finding new avenues of improvement.
- IF filters: Active and SAW filter technology may continue to improve. MEMS hold some hope for new filter technology.
- IF amplifiers: mature technology but some improvement in IP3 may occur.

7.11 Appendix: Power Level Received from an Adjacent Mobile Unit

As an example of how high power levels may be encountered at a receiver input, consider two mobile radios on cars located in adjacent parking spots. As a simplification we assume free space propagation and far field conditions. The frequency is 150 MHz. The antenna gain is 3 dBi. Transmitted power is 100 Watts. Distance between antennas is 4 meters.

$$f = 150 \cdot 10^{6} \cdot Hz$$

$$d = 4 \cdot m$$

$$G_1 = 2$$

$$G_2 = 2$$

$$P_t = 100 \cdot W$$

$$\lambda = \frac{300 \cdot 10^{6} \cdot \frac{m}{s}}{f} \qquad \lambda = 2\ m$$

$$P_r = P_t \cdot G_1 \cdot G_2 \cdot \left[\frac{\lambda^2}{(4 \cdot \pi \cdot d)^2} \right] \qquad P_r = 0.633\ W$$

REFERENCES

[1] A.D. Spaulding, F.G. Stewart, "An Updated Noise Model for Use in IONCAP," NTIA Report TR-87-212, U.S. Dept. of Commerce, Jan. 1987.

[2] S.M. Shajedul Hasan, S.W. Ellingson, "Optimum Noise Figure Specification," Chameleonic Radio Technical Memo No. 20, April 25, 2007. Submitted for publication in IEE Electronics Letters, review pending.

[3] H. A. Wheeler, "Fundamental limitations of small antennas," *Proc. IEEE*, vol. 69, pp. 1479 – 1484, Dec. 1947.

[4] T. Yang, W. A. Davis, W. L. Stutzman, and M.-C. Huynh, "Cellular Phone and Hearing Aid Interaction – An Antenna Solution," *IEEE Antennas and Propagation Magazine*, vol. 50, Issue 3, pp. 51-65, June, 2008.

[5] W. A. Davis, T. Yang, E. D. Caswell, and W. L. Stutzman, "Fundamental Limits on Antenna Size – A New Limit," submitted to *IEEE Transactions on Antennas and Propagation*, 2007.

[6] L. J. Chu, "Physical limitations on omni-directional antennas," *J. Appl. Phys.*, vol. 19, pp. 1163-1175, Dec. 1948.

[7] R.E. Collin and S. Rothschild, "Evaluation of antenna Q," *IEEE Trans. Antennas Propagation*, vol. AP-12, pp. 23-27, Jan. 1964.

[8] R. L. Fante, "Quality factor of general ideal antennas," *IEEE Trans. Antennas Propagation*, vol. AP-17, pp. 151-155, Mar. 1969.

[9] H. D. Foltz and J. S. McLean, "Limits on the radiation of electrically small antennas restricted to oblong bounding regions," *IEEE Antennas and Propagation. Symposium*, Orlando, July 1999.

[10] D.M. Grimes and C.A. Grimes, "Radiation of dipole generated fields," *Radio Science*, 34, pp. 281-296, 1999.

[11] A. D. Yaghjian and S. R. Best, "Impedance, bandwidth, and Q of antennas," *IEEE Trans. Antennas and Propagation,* vol. 53, Apr. 2005.

[12] T. Yang, W. A. Davis, and W. L. Stutzman, "Fundamental-Limit Perspectives on Ultra-wideband Antennas," submitted to *Radio Science*, 2007.

[13] M.W. Maxwell. *Reflections Transmission Lines and Antennas*, Newington CT: American Radio Relay League, 1990.

[14] R. A. Foster, "A Reactance Theorem," *Bell System Tech. J.* Vol-3, pp. 259-267**,** April 1924**.**

[15] S. E. Sussman-Fort, "Matching Network Design Using Non-Foster Impedances," *International Journal of RF and Microwave Computer-Aided Engineering*, Vol. 16, Issue 2, pp.135-142, Mar, 2006.

[16] R.C. Hansen, "Wideband Dipole Arrays Using Non-Foster Coupling," *Microwave and Optical Technology Letters,* Vol. 38, No. 6, pp. 453-455, Sep. 20, 2003.

[17] S. E. Sussman-Fort, Ron M. Rudish, "Non-Foster Impedance Matching for Transmit Applications," *2006 IEEE International Workshop on Antenna Technology Small Antennas and Novel Metamaterials,* March 6-8, 2006, p.p. 53 – 56.

[18] B. Brannon, (1998, Mar. 6). "Basics of Designing a Digital Radio Receiver (Radio 101)," Analog Devices, Inc., Greensboro, NC. [Online]. Available: http://www.analog.com/static/imported-files/tech_articles/480501640radio101.pdf

[19] Maxim (2002, Sep.). Application note 1197, Maxim Integrated Products, Inc. Sunnyvale, CA. [Online]. Available: http://www.maxim-ic.com/an1197

[20] U.L. Rohde, T.T. Bucher, *Communications Receivers Principles & Design* New York: McGraw-Hill, 1988, p. 260.

[21] Rohde, p.78.

[22] H.L. Krauss, C.W. Bostian, F.H. Raab, *Solid State Radio Engineering*, New York: Wiley, 1990, p. 267.

[23] Y.S. Poberezhskiy, "On Dynamic Range of Digital Receivers," *IEEE Aerospace Conference,* 3-10 March 2007, pp. 1-17.

[24] M. McHenry, et al., "Tuner Utilization and Feasibility (TUF) Study Final Report," Shared Spectrum Company. Vienna, Va. 2005).

[25] P. Baltus, R. Dekker. "Optimizing RF Front Ends for Low Power," *Proceedings of the IEEE,* Vol. 88, No. 10, pp. 1546-1559, Oct. 2000.

[26] P. Baltus. (2005, Aug. 2). "RF Transceiver Front Ends," Philips Semiconductors. [Online]. Available: www.w-i-c.org/events/2005/WIC%20midwinter%202005%20PB.ppt

[27] TeraVicta home page (2007, Dec.). TeraVicta Technologies, Austin, TX. [Online]. No longer available: http://www.teravicta.com

[28] RadiantMEMS home page (2008, Aug.). RadiantMEMS, inc., Stowe, MA. [Online]. Available: http://www.radantmems.com

[29] RadiantMEMS (2007, June 13). RMSW220D Data Sheet, RadiantMEMS, inc., Stowe, MA. [Online]. Available: http://www.radantmems.com/radantmems.data/Library/Radant-Datasheet220-NEW%20SPDT.pdf

[30] G.M. Rebeiz, *RF MEMS Theory, Design and Technology*, Hoboken, NJ: John Wiley and Sons, 2003, p. 192.

[31] L.H. Gemmer, "Electrical Breakdown between Close Electrodes in Air," *Journal of Applied Physics,* Vol. 30, No. 1, Jan. 1959, (Figure. 3.)

[32] A. Stehle, et al, "High-power handling capability of low complexity RF-MEMS switch in Ku-band," *Electronics Letters*, vol. 43, No. 24, 22 Nov. 2007.
[33] Rebeiz, p.45.
[34] J. Helszajn. *YIG Resonators and Filters.* New York: Wiley, 1985.
[35] S.T. He: "Research Progress in SAW Filter Banks" Journal of Zhejiang University Science. 2005 6A(9). Pp. 990-996.